Evolutionary analysis of mastrevirus functional regions

A thesis submitted in partial fulfilment of the requirements for the Degree of Master of Science in Biochemistry at the University of Canterbury.

By R.G. Lawry
University of Canterbury
2010

1 Literature Review 1

1.1 Introduction 1

1.2 The Geminiviruses : 1

1.3 The Mastreviruses 4

1.4 Molecular Biology of the Mastreviruses: 7

1.4.1 The Intergenic Regions 8

1.4.2 Proteins 9

1.4.3 The Replication Cycle and Viral Life Cycle 14

1.5 Viral Evolution 15

1.6 Modularity 19

1.7 Project Aims 21

2 Molecular Characterisation of a Novel Sugarcane Infecting Mastrevirus Species from South Africa 22

2.1 Abstract 22

2.2 Introduction 23

2.3 Methods 24

2.3.1 Sequencing and Isolation 24

2.3.2 Sequence Alignment 25

2.3.3 Recombination and Phylogenetics 25

2.4 Results 26

2.4.1 Genetic Structure 26

2.4.2 Evolution 45

2.5 Discussion 54

3 Selection Patterns and Modularity in the Mastrevirus Coat and Movement Proteins 57

3.1 Abstract 57

3.2 Introduction 58

3.3 Methods 64

3.4 Results 65

3.4.1 Sequence Conservation of Mastrevirus Coat Protein 65

3.4.2 Sequence Conservation of Mastrevirus Movement Protein 71

3.4.3 MSV-Kom and MSV-Set Comparison 71

3.4.4 Conservation by Host species 73

3.4.5 Selection 77

3.4.6 Selection by Host 83

3.4.7 Codon Usage 89

3.5 Discussion 91

3.5.1 Sequence alignment and Conservation 91

3.5.2 Selection 92

3.5.3 Selection by Host 95

3.5.4 Codon Usage 95
3.5.5 Implications of Selection and Conservation Patterns 96

4 Selection and conservation patterns in the mastrevirus replication protein..... 97

4.1 Abstract 97

4.2 Introduction 98

4.3 Methods 101

4.4 Results 102
4.4.1 Selection 109
4.4.2 Selection by Host 114
4.4.3 Codon Usage 118

4.5 Discussion 119
4.5.1 Sequence alignment and Conservation 119
4.5.2 Selection 121
4.5.3 Selection by Host 121
4.5.4 Codon Usage 122
4.5.5 Implications of Selection and Conservation Patterns 122

5 Conclusion 124

6 References 129

Acknowledgements

It is an honor for me to thank those who made this thesis possible.

I would like to particularly thank my supervisor, Arvind Varsani, who provided excellent advice, stimulated my interest in a new field of science and who encouraged me in my studies.

I am grateful for the support of my parents, Dave and Shirley Lawry, who provided support throughout the time I was writing the thesis.

Lastly, I would like to dedicate this thesis to my grandmother, Margaret Dorman, who has supported me throughout the years with good conversation and sage advice.

Abbreviations

AfSV – African streak viruses
BeYDV – *Bean yellow dwarf virus*
bp - Base pairs
CP – Coat protein
CpCDPKV - *Chickpea chlorotic dwarf Pakistan virus*
CpCDSDV - *Chickpea chlorotic dwarf Sudan virus*
DNA – Deoxyribonucleic acid
dNTP – deoxynucleoside triphosphates
dsDNA – Double stranded DNA
DSV – *Digitaria streak virus*
ESV – *Eragrostis streak virus*
FEL – Fixed effects likelihood
GRAB – Geminivirus RepA binding
ICTV – International committee on Taxonomy of Viruses
IFEL – Internal fixed effects likelihood
kDa - Kilodaltons
LIR – Long intergenic region
MiSV – *Miscanthus streak virus*
ML – Maximum likelihood
MP – Movement protein
MSV – *Maize streak virus*
NAC - NAM (No apical meristem), **A**TAF, **C**UC2 (Cup-shaped cotyledon 2)
NG - Nigeria
NLS – Nuclear localisation signal
ORF – Open reading frame
PanSV – *Panicum streak virus*
pRBR – Plant retinoblastoma related protein
RBR - Retinoblastoma related protein
RCR – Rolling circle replication
REL – Random effects likelihood
Rep - replication protein
RepA – Replication associated protein
RNA – Ribonucleic acid
SacSV – *Saccharum streak virus*
SIR – Short intergenic region
SISV – Sugarcane infecting streak virus
SLAC – Single likelihood ancestor counting
SSEV – *Sugarcane streak Egypt virus*
SSRV – *Sugarcane streak reunion virus*
SSV – *Sugarcane streak virus*
ssDNA – Single stranded DNA
TbyDV – *Tobacco yellow dwarf virus*
tRNA – Transfer RNA

TYLCV - *Tomato yellow leaf curl virus*
USV – *Urochloa streak virus*
WDV - *Wheat dwarf virus*
3D – 3 dimensional
RDP3 – Recombination detection program 3

1 Literature Review

1.1 Introduction

New and emerging virus species are becoming an increasing threat to our way of life economically and physically. Plant viruses are particularly significant as they affect our food supply and are capable of rapidly spreading to new plant species. *Geminiviruses* are a group of viruses that highlight this phenomenon well. Indeed Geminiviruses are some of the earliest recorded plant viruses being described as far back as 752 AD in a Japanese poem written to describe geminivirus symptoms in eupatorium leaves (Saunders *et al.,* 2003). More recently, and in a more threatening manner, *Geminiviruses* have adapted to infect key crop species such as maize, sugarcane, tomatoes, beet and many more.. An example of this is the introduction of grasses such as Maize into Africa, which allowed a species jump for mastreviruses, which were endemic in native grasses (Varsani *et al.,* 2008a). Over a relatively short period of evolutionary time a number of new *Geminiviruses* have emerged, making them a good model for understanding the evolution and spread of new plant pathogens. The economic importance of *Geminiviruses* also makes an understanding of their mechanisms of adaptation crucial in preventing new emergence and minimising the impact of current strains.

1.2 The Geminiviruses :

Geminiviruses are a group of single stranded DNA (ssDNA) viruses that infect a number of plant species. Geminiviruses are identifiable by a number of unique characteristics. One of the first to be identified was the geminate (or twinned) quasi-icosohedral capsule, made up of 22 capsomers and around 22 by 38nm in size, that surround the virus (Hatta & Francki, 1978, Zhang *et al.,* 2001). The name *Geminivirus* decends from this observation. Other distinguishing features of *Geminiviruses* are the circular ssDNA genome (Harrison *et al*., 1977), which is usually between 2700-3000bp in length and the TAATATTAC nucleotide sequence which is the origin of viral rolling circle replication.

Mastreviruses symptoms range from leaf curl to mosaic patterns to streaks. The symptoms of Maize streak virus are shown below (Figure 1.1) (Fuller 1901, Shepherd *et al*., 2010).

Figure 1.1: Characteristic streak patterns from on a Maize leaf infected by Maize streak virus. Photo Courtesy of Frederik Kloppers. (Source Shepherd et al., 2010)

The family *Geminiviridae* is currently composed of 4 genera which are characterised by their vector, host specificity, genome arrangement and host range. There is also significant phylogenetic evidence supporting each of these genera. These genera are the *Begomovirus* (*e.g.Bean golden mosaic virus;*Galvez & Castano, 1976, Howarth *et al.,* 1985)*, Topocuvirus* (*e.g Tomato pseudo curly top virus;* Giddings *et al*, 1951)*, Curtovirus* (*e.g.Beet curly top virus;* Ball, 1909, Stanley *et al*., 1986)*, and Mastrevirus* (*e.g. Maize streak virus;* Fuller, 1901, Storey *et al* 1925, Howell *et al*., 1984). Additionally, two viruses have recently been characterised which do not readily fit into the present taxonomic groupings and may ultimately form the founding members of two new genera – Beet curly top Iran virus (BCTIV; Yazdi *et al*, 2008)) and Eragrostis curvula streak virus (ECSV; Varsani *et al,* 2009).

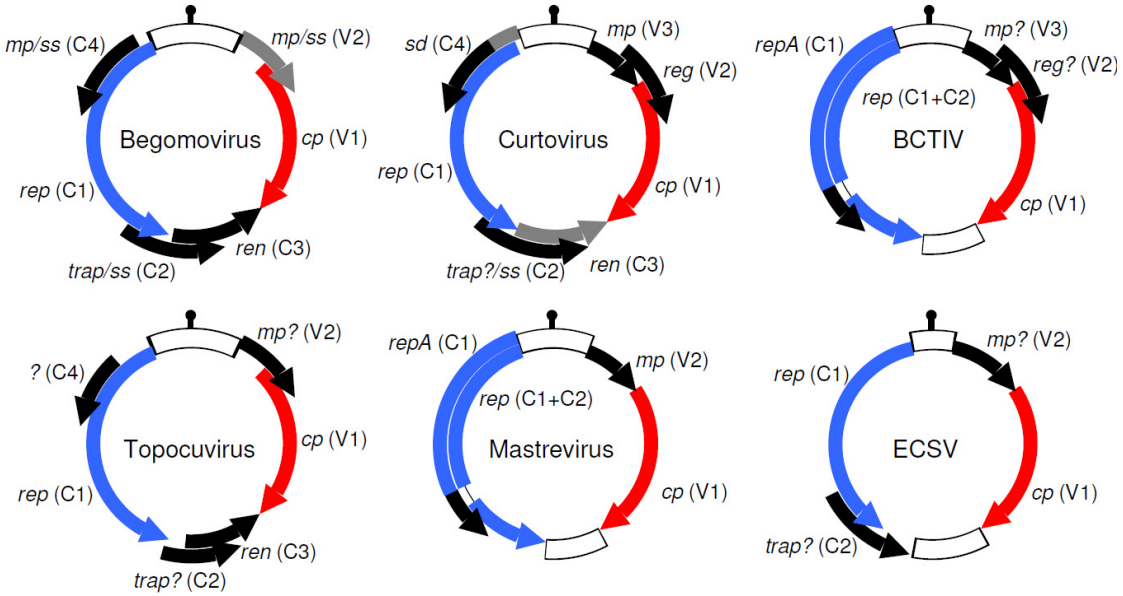


Figure 1.2: The genomic structure of the geminiviruses (Note only DNA A of Begomoviruses is shown). Open reading frames and their direction of reading are shown by each arrow. V stands for virion sense, C for complementary sense. Proteins encoded by these are: Movement protein (MP), Coat protein (CP), replication protein (Rep/A), replication enhancer protein (ren), transcription activator protein (trap), silencing suppressor (ss) symptom determinant (sd) and a regulatory protein (reg). In mastreviruses Rep can be spliced to produce two different transcripts. It is suspected that BCTIV rep also undergos splicing. Non coding regions are shown by a curved box. The conserved stem loop structure (TAATATTAC) is shown by the black line in the long noncoding region known as the Long intergenic region in *mastreviruses*. In Begomoviruses this Is known as the common region, and in Topocuviruses and Curtoviruses it is the Intergenic region. Mastreviruses posess a second intergenic region known as the Short intergenic region. ECSV has unconventional noncoding regions, and they have been labeled IR-1 and 2 (intergenic region 1 and 2) (Source Varsani *et al*., 2009, Used with permission)

Begomoviruses are the largest group of *Geminiviruses* with over 100 unique species (Fauquet & Stanley., 2005). They are generally distinguishable by their bipartite genomes, transmission vector and host species. However, there are a few monopartitie begomoviruses such as *tomato yellow leaf curl virus* (TYLCV). A unique feature of the begomoviruses is their separately encapsulated bipartite genome. The genome sections are called DNA A and DNA B and encode different proteins. The genome is more complex then several other genera, encoding for a different and larger group of proteins. Begomoviruses have a highly specific vector, being transmitted only by a single species of whitefly (*Bemisia tabaci).* Begomoviruses also have been found to only infect dicotyledonous plants.

Curtoviruses differ from the Begomoviruses in that they have a monopartite genome with a different arrangement (Figure 1.2) and are transmitted by leafhoppers. Topocuviruses are one of the smallest groups of *Geminiviruses* and have an intermediate mix of traits, possessing a monopartite genome, similar to that of the Curtoviruses, but are instead transmitted by treehoppers (Fauquet & Stanley, 2003).

Beet curly top Iran virus has a curto-like virion sense genome but a mastre-like complementary sense genome with a replication associated gene (Figure 1.2). Eragrostis curvula streak virus is the most recently discovered *Geminivirus* which has unusual genomic features. It has a Begomovirus like transcription activator protein and a Mastrevirus like coat protein (Varsani *et al*., 2009). These are particularly unusual as the mix of genes appears not to originate from recombination, making this virus a good identifier of an ancestral geminiviral state.

The final genus is the Mastreviruses. These viruses infect a wide variety of wheat and maize species and have the simplest genome of the *Geminiviruses*. These are the primary focus of this study and will be described in greater detail below.

1.3 The Mastreviruses

The Mastreviruses are diverse genera of *Geminiviruses* which infect both mono- and dicotyledonous plants. Currently there are 20 known species of Mastrevirus (GenBank; http://www.ncbi.nlm.nih.gov/taxonomy). The majority of these are from the African streak viruses (AfSV) group, comprised of Maize streak virus (MSV; Howell *et al*., 1984), Sugarcane streak virus (SSV; Hughes *et al.,* 1993), Sugarcane streak reunion virus (SSRV; Shepherd *et al.,* 2008), Sugarcane streak Egypt virus (SSEV; Biggare *et al.,*1999*)*, Panicum streak virus (PanSV; Briddon *et al.,* 1992), Urochloa streak virus (USV; Oluwafemi *et al.,* 2008*)* and Eragrostis streak virus (ESV; Shepherd *et al.,* 2008). The phylogenetic relationships between the AfSVs are shown below (Figure 1. 3).

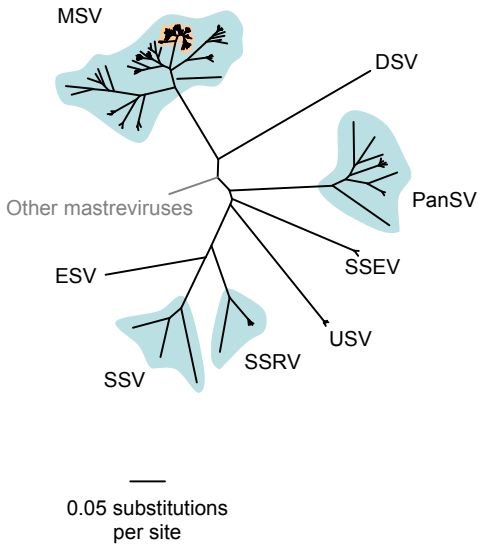


Figure 1.3: Phylogenetic relationships between various AfSV-like mastrevirus species. (Source Shephard *et al*, 2010)

Mastreviruses that infect monocotyledonous hosts have predominantly been found in Africa, with the exception of *Miscanthus streak virus* (MiSV; Chantani *et al., 1991*) from Japan, *Digitaria streak virus* (DSV; Donson *et al*., 1987) from Vanuatu, *Wheat dwarf virus*, *Oat dwarf virus* and *Barley dwarf virus* originating from Europe, Asia and the Middle East (MacDowell *et al*,. 1985, Köklü et al., 2007, Schubert et al., 2007, Woolsten et al., 1988, Xie et al., 2007), and *Chloris striate mosaic virus* (CSMV; Andersen 1998) from Australia. Within each of these species are a number of strains, each of which have varying degrees of virulence and have a different primary host species. Dicotyledonous infecting master viruses include *Bean yellow dwarf virus* (BeYDV; Liu *et al*., 1997b), *Tobacco yellow dwarf virus* (TbYDV; Morris et al., 1992) and chickpea chlorotic dwarf Sudan virus and chickpea chlorotic dwarf Pakistan virus (CpCDSDV, CpCDPKV) (Nahid et al., 2008; Makkouk et al., 1995).

Mastreviruses are the causal agents of disease in many important crops. WDV, MSV and the sugarcane streak viruses all have significant impacts on crops. For example MSV infections routinely reduce crop yields by over 70% in susceptible varieties (Bosque-Perez *et al*., 1998). The extent of crop damage has been closely linked to the time of infection in a plants life cycle, with younger plants suffering proportionately greater

damage. Farming techniques, the density of maize distribution and vector distribution all greatly influence the development and longetivity of infections, and can provide ways of mitigating the damage caused by Mastreviruses (Bosque-Perez., 2000, Shepherd *et al.,* 2010).

Due to maize being a primary food source in Africa, MSV has been extensively studied, and is perhaps the most comprehensively understood of all the Mastreviruses. There are currently 11 known strains of MSV (Figure 1.4), of which MSV-A is the most significant in causing maize streak disease. Other strains can infect maize, but they are far less virulent and cause significantly less damage (Martin *et al.,* 2001). Recent studies have shown that they may be capable of damaging other crops such as wheat, and so their effects upon crops cannot be entirely discounted.

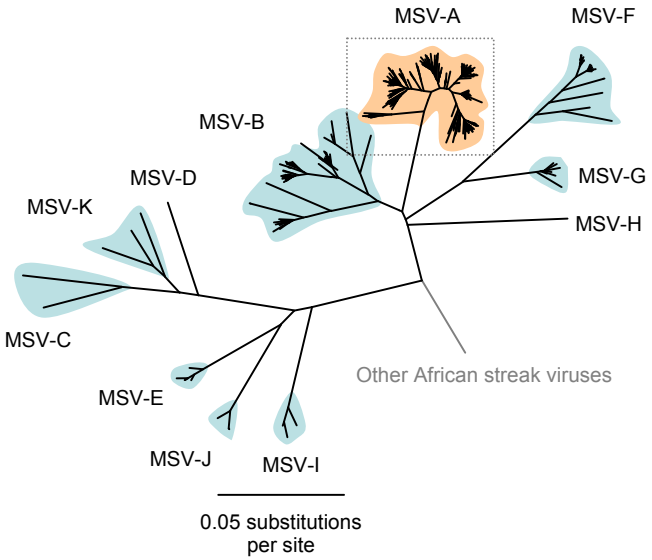


Figure 1.4: Phylogenetic relationships between the 11 known MSV strains. Only MSV-A causes Maize streak disease. MSV-B and C are known to cause disease symptoms in, wheat and barley, and have an as yet unknown but possibly significant effect on agriculture. Other strains are generally found infecting wild grass species. (Source Shephard *et al*, 2010)

1.4 Molecular Biology of the Mastreviruses:

The Mastreviruses are distinguished from other *Geminiviruses* by their monopartite genome, transmission vector (*Cicadulina* leafhopper species). Most of them infect monocotyledonous plants with the exception of BeYD, TbYDV, CpCDSDV and CpCDPKV which infect dicotyledonous plants. Mastrevirus particles contain an 80 nucleotide (nt) sequence, that acts as a primer, bound to the ~2.7kb ssDNA genome. They also undergo a splicing event on the complementary sense transcript to yield a functional replication associated protein. As MSV is the most well characterised of the Mastreviruses, it provides an excellent basis for understanding the molecular mechanics of the group.

Mastrevirus genomes are between 2500-3000 nucleotides long, and consist of circular single stranded DNA (ssDNA) (Lazarowitz *et al.,* 1989, Howell *et al*., 1985). There are four coding and two noncoding regions on the genome, which are read bidirectionally, commonly referred to as the virion (V) sense open reading frames (ORFs) and complementary (C) sense ORFs (Morris-Krsinich *et al*., 1985). The coding regions code for four different proteins, coat protein (CP), movement protein (MP), replication associated protein (RepA) and full length replication protein (Rep) (Lazarowitz *et al*., 1989). Each direction contains information for two of the coding regions, with the noncoding regions containing regulatory elements such as promoters and enhancers, as well as the replication origins. Generally regions in the V sense are involved in movement and structure, and those on the C sense are associated with replication and transcription. Figure 1.5 shows an overview of the mastrevirus genome.

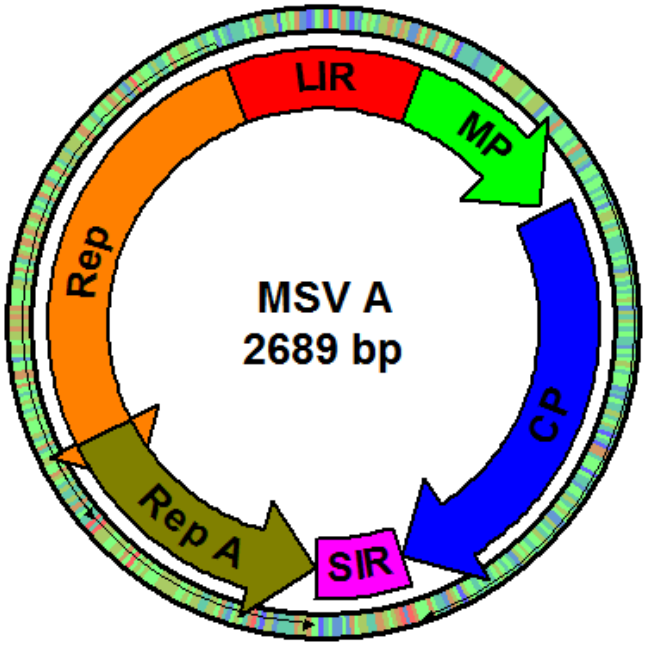


Figure 1.5: The layout of the Mastrevirus genome, showing different directions of reading. Note the overlap in Rep and RepA, showing the intron. Colors represent GC nucleotide richness in each area, with blue being the highest and red being the lowest GC level.

1.4.1 The Intergenic Regions

There are two intergenic regions on the mastrevirus genome, the long intergenic region (LIR), and the short intergenic region (SIR). In MSV the LIR is around 300 nt long and the SIR is around 150 nt long. Both are responsible for regulation of gene expression and provide protein binding sites and replication origins for the opposite directions of replication. The LIR and SIR are both highly variable sequences, and tend to show the highest diversity between strains.

The LIR

The LIR contains the V sense origin of replication, the TAATATTAC sequence, which is one of the most highly conserved sequences in *Geminiviruses* (Heyraud *et al*., 1993, Lazarowitz 1987). This sequence is present as a stem loop structure and is responsible for the initiation and termination of replication and the efficiency at which replication occurs (Fenoll *et al*., 1988). While the TAATATTAC sequence is highly conserved, with even minor variations causing loss of replication ability, the stem of the structure is reasonably

tolerant to variation, especially in mastreviruses, with different sites along the stem having varying effects on replication efficiency. An area of particular interest is the so called upstream activator sequence. This region contains G-C rich regions and promoter elements that are both responsible and essential for activation of transcription in MSV (Fenoll *et al*., 1990).

The SIR

The SIR contains the C sense origin of replication, and is often associated with a short primer oligonucleotide sequence, that allows attachment of replication machinery (Fenoll *et al*., 1988). It is essential for the activation of the negative strand replication mechanism. The SIR contains groups of polyadenylation signals and it appears to influence transcription and replication signals, particularly those to do with replication termination. SIR is one of the most tolerant regions of the genome to variation.

1.4.2 Proteins

Movement Protein

The MP is a 10.9 KDa protein that is generally around 110 amino acids long (Boulton *et al*., 1989). It contains one currently identified region between positions 38-63 which has a hydrophobic potential trans-membrane domain. MP is responsible for regulation of the coat protein, cell to cell movement of viral particles, cell localisation and may have some effects on the toxicity of the virus to the host plant. Introns of the movement protein transcript have been shown to affect the levels of CP expression. This may be crucial for infectivity due to the effects of CP-MP interactions on viral movement (Liu *et al.,* 2001). MP has been shown to bind CP allowing control of viral positioning in the cell. It has been shown to localize towards the cell periphery especially around plasmodesmata which provide networks between plant cells, further supporting its role in inter-cell movement (Dickenson *et al*., 1996). The movement protein has inherent toxicity, as demonstrated by the stunting of transgenic MP expressing plants (Hou *et al*., 2000), and variations in the MP appear to be responsible for some of the variation in toxicity between viral strains. It is currently unclear whether the toxicity is due to actual increases in MP toxicity or to increased concentration of MP in more virulent strains.

Coat protein

The coat protein occurs naturally as a 28KDa protein. Each protein chain is approximately 250 amino acids long. CP is responsible for transport of viral DNA, localisation of the virus within the host and for vector specificity. It is essential for replicative function of the virus, as well as its ability to cause systematic infections (Boulton *et al*., 1989). The structure of the MSV-A [NG] coat as a whole has been solved (Figure 1.6) and consists of 22 capsomers arranged in a quasi icosohedral shape with pentomeric symmetry. Each head of the virus coat is composed of 11 of these capsomers, each of which is composed of 5 CP structures. Models of the individual CP structure have also been completed and more clearly illustrate the 5 fold symmetry of the molecule.

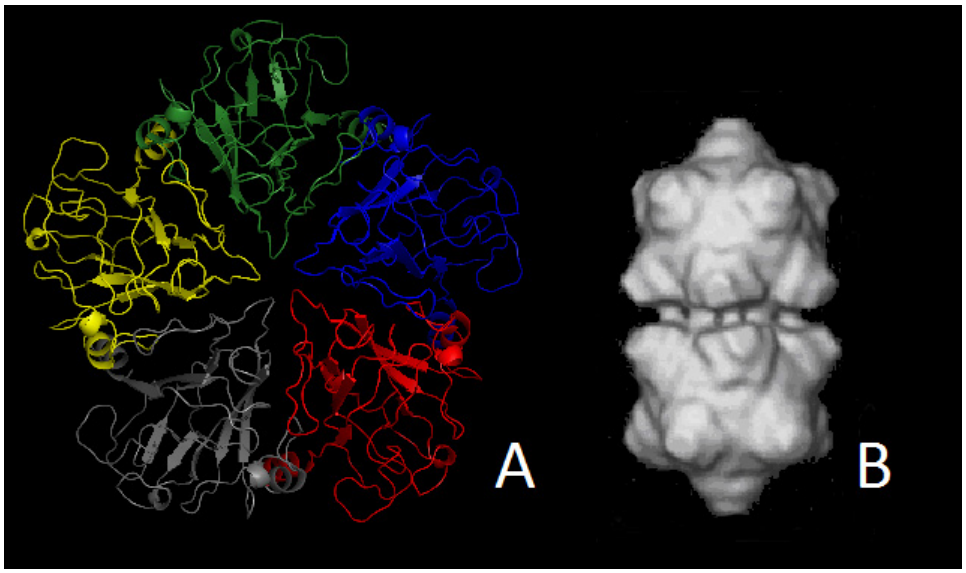


Figure 1.6: 3D Structures of: A) a capsomer drawn in cartoon form, showing individual CP subunits in colour. The 5 fold symmetry is clearly visible. B) Structure of the full MSV-A[NG] capsid (Zhang et al., 2001). The full capsid is 22nm in diameter. Note the two heads, each containing 11 capsomers. Image B taken from Zhang et al (2001).

CP has been show to bind ss and dsDNA non specifically which may facilitate its ability to localize to the host cell nucleus, as well as allowing organized construction of viral particles (Liu *et al.,* 1997a, 1999). Two key regions of the protein have been identified in MSV, the first being a nuclear localisation signal between positions 9-22. This region contains numerous basic residues, which is characteristic of DNA binding motifs. It is also notable that the entire CP is highly basic, with a pK_a of 10.4 (Liu *et al.,* 1997a). This

may influence binding to DNA but is unlikely to be the sole factor influencing CP function. The second region has been demonstrated as the key binding region for DNA, being a 104 amino acid stretch originating at the N terminus of the protein (Liu *et* al., 1997a). The C terminal sequence appears to be unnecessary for DNA binding. Nuclear localization is thought to occur by binding of CP to viral ssDNA, although the exact mechanism is currently unknown.

CP is also thought to interact with MP in order to facilitate cell to cell movement and localisation within the cell (Kotlizky *et al*., 2000). This is thought to occur in a manner similar to that of begomovirus proteins, although exact details of the interactions in mastreviruses are as yet unknown. Experiments with chimeric viruses have shown that coat protein is the key determinant of vector specificity (Briddon *et al*., 1990). When genes for CP from two different viruses were exchanged, each virus became transmissible by the vector of the other. The exact mechanism behind this specificity is currently unknown.

Rep/RepA

The replication associated proteins Rep and RepA are key determinants of the replication function in Mastreviruses. Both proteins are very similar, sharing a large overlapping area which serves as an intron for Rep splicing (Shalk *et al*., 1989). The Rep protein is approximately 360 amino acids in length and contains numerous key binding sites. Rep A is a shorter protein, around 150 amino acids long, being identical to Rep on the N terminus until the splice site. Currently only one part of Rep's 3D structure has been determined, and only in Tomato yellow leaf curl virus. This helicase domain is shown below.

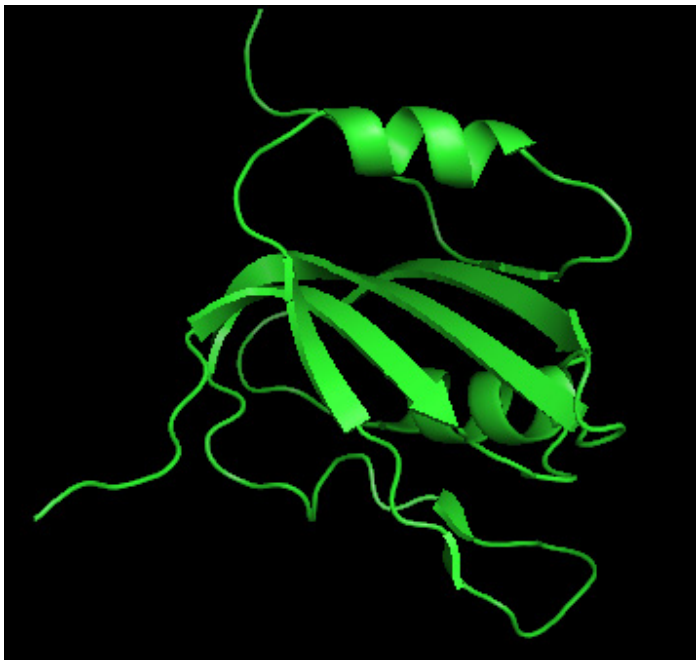


Figure 1.7: TYLCV helicase domain showing two alpha helices and the central beta sheet.

RepA contains three known binding sites, allowing interaction with a number of non viral proteins. It is also known to be capable of interacting with itself, forming oligomers that can interact with viral DNA, with an as yet unknown function. RepA contains an **L**x**C**x**E** (Leucine-x-Cysteine-x-Glutamic acid, where x is variable) amino acid motif, which allows binding to proteins of the Retinoblastoma family (pRBR, RB, RBR) (Xie *et al*., 1995). Full length Rep is incapable of binding this family, possibly due to the extended length of the protein interfering with molecular interactions. The C - terminal of RepA contains a NAC (**N**o apical meristem [**N**AM], **A**TAF 1, **C**up-shaped cotyledon 2 [**C**UC2] genes) domain which is able to bind various plant proteins associated with plant growth and development (Gutierrez., 2000). It is also able to bind Geminivirus Rep-A binding (GRAB) proteins which may influence the ability of geminiviruses to influence plant growth and development (Xie *et al*., 1999). Only the final 37 C-terminal amino acids of RepA are necessary for these proteins to bind. The C - terminus of RepA has been shown to have transactivation ability (the ability to increase the rate of gene expression) on yeast growth and elongation factors.

Full length Rep is the only protein that is absolutely required for DNA replication in Mastreviruses (Gutierrez., 2000). It is responsible for the initiation of replication during rolling circle replication (RCR), has DNA and NTP binding ability, ATPase ability and transactivation ability. Rep can initiate replication due to three key motifs, known as RCR I – III (Ilyina and Koonin., 1992, Gutierrez., 2000,). RCR III (**V**x**DY**xx**K**) is responsible for nicking of the TAATATT/AC sequence (/ indicates nick site), for which the **Y** and **K** residues are crucial. Initiation is then further determined by iterons that flank the TAATATTAC sequence. These are thought to interact with RCR I in an as yet undetermined manner. RCR II is believed to have metal chelating ability for divalent cations such as $Mg^{2}+$ which are thought to be crucial for activity. These motifs are all present within the 120 N terminal amino acids, which are also crucial for DNA binding activity.

The C terminal domain of Rep contains NTP binding function as well as ATPase activity (Gorbalenya and Koonin, 1989). This activity is conferred by the presence of Walker A and B motifs, which are commonly found in ATPases. However the exact purpose of this activity in mastreviruses is unknown, as it seems to be unnecessary for Rep activity. The C - terminus also contains a transactivation domain, which may allow Rep to regulate viral gene activity (Horvath et al., 1998).

Recent research indicates that the 3D structure of the Rep DNA may have a significant role in viral infectivity. A mutant strain of MSV with a 3 nucleotide substitution in Rep's pRBR binding motif was shown to have no pRBR binding capability and slightly reduced infectivity in maize, although viral replication could still occur. However a single nucleotide reversion from C (601) to A occurred in such viruses with very high frequency, eventually entirely supplanting the original mutant population (Shepherd *et al*., 2005). Intriguingly this reversion had no apparent effect on restoring retinoblastoma related binding protein binding ability, and did return the virus to the same infectivity as the wild type. Significantly the C to A reversion is a low probability substitution in mastreviruses, indicating the existence of another selective force acting upon this site.

Structural modelling using RNA algorithms showed significant secondary structure changes in the C(601) mutant, whereas the A(601) reversion closely resembled the wild type virus. Temperature dependent ultraviolet absorbance assays showed similar results, with the mutant again being noticeably different in profile from wild type and reversion. The results demonstrated a strong selective advantage for mutants that maintain the genomic secondary structure whilst ruling out the possibility of protein level changes having an effect (Shepherd *et al*., 2005). This provides evidence that for Rep at least, maintenance of DNA structure may be as necessary for mastrevirus function as the proteins the DNA encodes. While the exact mechanisms of this are currently unknown, factors such as enhanced splicing, transcription or even reduced susceptibility to attack by host defence mechanisms may be instrumental in providing the strong selection that is in evidence. Other studies in begmoviruses have seen similar effects occurring, suggesting that DNA secondary structure effects may be an important and widespread influence on mastrevirus evolution.

1.4.3 The Replication Cycle and Viral Life Cycle

The life cycle of the mastrevirus begins when it is injected into a plant host cell by a *Cicadulina* leafhopper. As little as five minutes of feeding by a leafhopper is enough for this to occur. Once within the cell the viral coat is removed and viral DNA is localised to the plant nucleus predominantly by the CP, although MP may be involved (Liu *et al*., 1997a, 1999).

Once within the nucleus DNA replication may occur (Figure 1.8). Geminiviral DNA replication occurs by a rolling circle mechanism, which has three distinct stages. During the first stage of RCR the structure of the viral DNA changes from the normal circular form to a supercoiled double stranded circular form (known as Replicative Form I). This form allows stage two to initiate, during which expression of the replication proteins begins. Rep can then bind to the TAATATTAC stem loops structure of RFI allowing replication to begin. Once the new strand of DNA is synthesised step three of replication may occur. In this stage viral DNA is encapsidated and new viral particles are formed.

These are then localized to other cells by MP interactions (Boulton *et al*., 1989), where the lifecycle can continue. When a leafhopper feeds upon a plant infected with MSV for a sufficient time period, viral particles will be transmitted to the insect. Viral particles accumulate in the gut, and can then be passed on to nearby plants.

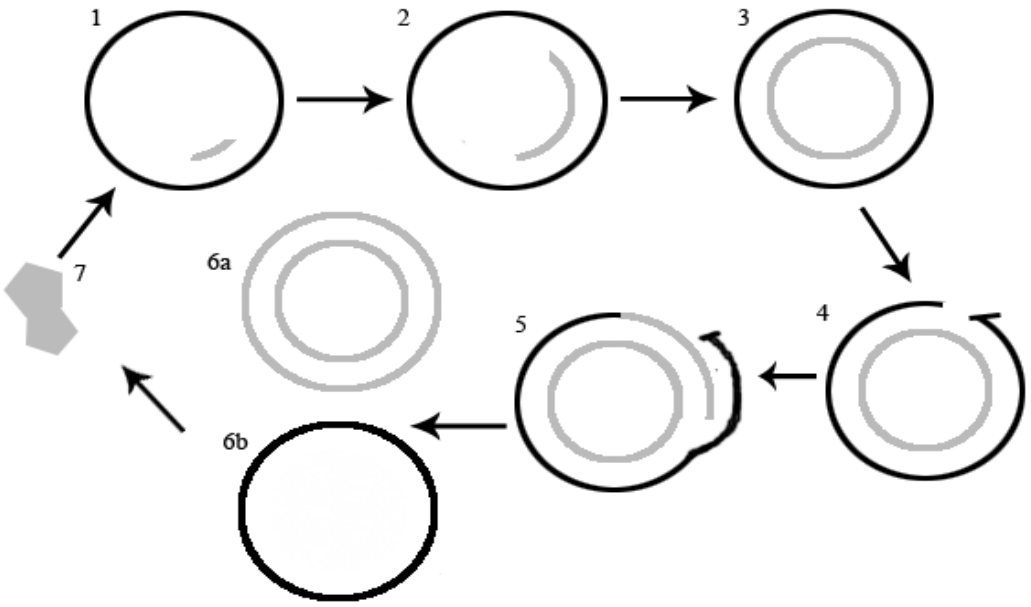


Figure 1.8: The replication cycle of a mastrevirus. 1) Viral DNA in its unencapsidated form. Note the short primer attached to the SIR. 2) Primer elongation begins, replicating the – sense strand. 3) Replicative form II is achieved when the – strand replication is complete. The DNA then forms double stranded supercoils. 4) Nicking opens up the + strand for replication. 5) Replication proceeds. 6a and b) Two new copies are produced. One is still double stranded and will go through further RCR cycles. The other is encapsidated into a viral particle (7).

1.5 Viral Evolution

DNA viruses have traditionally been regarded as relatively slowly evolving viruses, especially when compared to RNA viruses. Recently however evidence has emerged that some ssDNA viruses such as nanoviruses can evolve as swiftly as RNA viruses (Duffy *et al*., 2008). Much evidence, including the recent dating of the origin of mastreviruses is supporting the view that Geminiviruses are capable of very swift evolution, due to a variety of mechanisms. The two main reasons for the ability of Geminiviruses to rapidly evolve are relatively high mutation rates and recombination.

The high mutability at the nucleotide level has a number of probable causes, a number of which are related to the single stranded nature of the geminiviral genome. ssDNA is by nature less stable then double stranded, due to a lack of complementary base pairing. The inability of numerous repair methods to operate on ssDNA also makes it more prone to mutation. As viruses use the host's DNA replication mechanisms they often have no intrinsic DNA repair facility. This means they are reliant on host DNA repair capability which is often ill suited to action upon the viral genome. In the case of Geminiviruses it is theorised that excision repair systems cannot function upon the ssDNA of the geminiviral genome as there is no complementary strand to use as a template, or that the complementary strand is present for too short a time for these systems to effectively operate (Duffy & Holmes, 2008). Systems such as mismatch repair may also be unable to act upon Geminiviruses due to low levels of methylation. Furthermore host defense processes may actively target viral DNA. Plant systems frequently attack viruses by oxidation, and it is possible that ssDNA is more vulnerable to oxidation then dsDNA (van der Walt *et al*., 2008). Experimental evidence supporting this is that the level of G to T transversions is high, which is frequently seen upon oxidation of guanine. Secondly, it has been observed that mutations affecting DNA secondary structures are rapidly selected against (Lefeuvre *et al*., 2007a). This may be due to protective effects of secondary structure against host defence mechanisms.

As well as the numerous effects mentioned above, it is known that mastrevirus basal mutation rates are up to or in excess of $2x10^{-4}$ mutations per site per year (Harkins *et al*., 2009a, van der Walt *et al*., 2008). This is exceptional for a DNA virus and comparable to many RNA viruses. However, while this rate allows for generation of significant novelty it does not account for the effects of positive and purifying selection. The substitution rate gives a more accurate estimate of true evolutionary occurrences and current evidence places it in the range of 10^{-3} and 10^{-5} substitutions per site per year accumulating in the general population (van der Walt *et al*., 2008). Surprisingly data gathered from both lab and field experiments with begomovirues support rates around this level, potentially indicating that selection pressures are not strongly affecting substitution rates. The high rates of substitution and lack of evidence for strong selection pressures on at least some

sections of geminiviral genomes indicates that rapid accumulation of novel neutral mutations is possible in mastreviruses. It also suggests that recent theories of geminiviral co-divergance with plant hosts are less likely, as they require far lower substitution rates, and much stronger selection then is so far in evidence (Harkins *et* al., 2009b, Wu *et al*., 2008)

Determinaton of a reasonable estimate of basal substitution rates has been important to establishing a timeline of mastrevirus evolution. A recent experiment performed by Harkins *et al* (2009b) estimated a basal genome wide substitution rate of MSV-A type mastrevirues as $3.5x10^{-4}$ a figure comparable to that for geminivruses in general. Using this data and a relaxed molecular clock model the common ancestor of MSV-A was placed between 1839 and 1905, a figure that agrees with field observations from that time (Fuller, 1901). A more accurate estimate was made possible by the fact that CP of MSV-A originates from MSV-B and is highly homologous to it. This meant that using the basal substitution rate within CP and comparing it to the MSV-B origin allowed determination of MSV-A/B last common recombination event to be in the 19th century. This pointed to MSV-A and B having a common ancestor in the middle of the 18th century (Harkins *et al*., 2009b). The significance of these figures is made clear when it is know that maize was only introduced into Africa in the 1500s, and that Maize Streak Disease was noted in 1870. These dates indicate that MSV-A took a maximum of 130 year and a minimum of 20 years to become highly virulent and widespread in Africa.

Recombination is the second driving force behind the rapid evolution of mastreviruses. Recombination is very well documented in this group, particularly in MSV itself, in which the major crop infecting strain, MSV-A, is a recombinant between two more innocuous grass infecting strains, a recombinant of MSV F / G and MSV B (Varsani *et al.,* 2008a). Recombination is crucial in viruses as it in some respects mimics the benefits of sexual reproduction. Accumulation of deleterious mutations can be avoided, and diversity of positive mutations can be maintained throughout the species. Recombination also allows generation of novel variation by a mixture of segments of a genome, although this is constrained by various genetic mechanisms.

Mastreviruses, and indeed Geminiviruses in general, show a strong capability for recombination. Recent studies have shown that ssDNA viruses in general have a high propensity for recombination, being able to swap genomic regions both between and within species. Mastreviruses however have predominantly been observed to recombine within species. This was made particularly clear in a study by Varsani *et al* (2008a) in which it was observed that only 4 of 11 major MSV strains had crossovers covering less then 30% of the genome. Other studies have shown similar patterns in PanSV and the SSVs. The evolutionary advantage of recombination in mastreviruses was clearly documented by van der Walt *et al* (2009) where it was shown that not only do strains of MSV recombine frequently, they rapidly produce significantly more virulent varieties of the virus over fairly short periods of time. This is again supported in nature by MSV-A being the most virulent variety in maize (Martin *et al*., 1999).

The exact reason for the high recombination in mastreviruses is under debate, however several theories have been put forward. Recombination within Mastreviruses has two definitive 'hotspots' and one coldspot. The hotspots are present in the V- origin of replication and around the interface between CP and the SIR (Varsani *et al*., 2008a). These hotspots, particularly that around the CP/SIR interface, may indicate clashes between replicative mechanisms and transcription mechanisms (Lefeuvre *et al*., 2007b). It has also been noted that the primer for replication binds near the hotspot in the SIR, which may have further effects upon the stability of this region. It has also been suggested that the SIR and C-terminal domain of the CP are relatively tolerant to the effects of recombination, and this allows a higher frequency of recombination to occur in this region. DNA repair mechanisms of the host plant may allow for high levels of recombination, as the host may not select correct template strands for DNA repair, thus inserting numerous recombinant sections into a host genome dependent on which viral DNA strand is used in a template. This may also explain the high levels of recombination that can occur in a single replication cycle.

1.6 Modularity

Mastreviruses are ideal for studying the modularity of genes due to their small number of genes with complex interactions. A modular set of gene is one in which the full range of a genes function can occur without external interactions. Sets of genes are often referred to as gene cassettes. Thus the SIR is a module, as it functions well by itself in a foreign genetic environment. The CP-MP cassette is also a module, as discussed later. Modularity has significant effects on evolution as it creates selection pressures on areas of genes that may at first appear to be non-functional regions. This occurs as the environment within a module may be important its function. Disruption of this environment would thus be selected against. Mutations that act to maintain or enhance interactivity of genes within a module become favoured as they increase the efficiency of that module. Quite apart from the evolutionary effects modularity allows a small number of genes,that may have a single function alone, to perform numerous functions by interaction with other genes. This is crucial in mastreviruses as the limited size of the genome means that all necessary functions must be performed by only three genes.

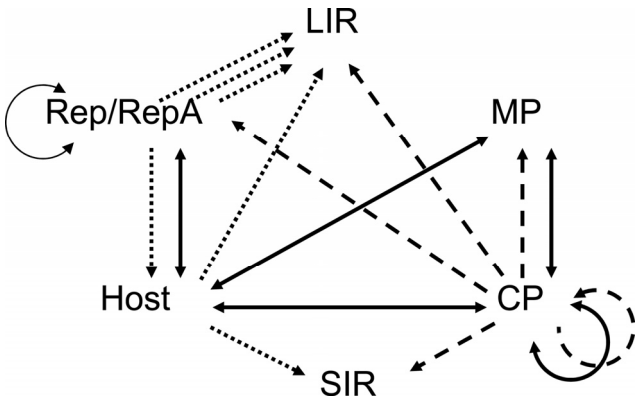


Figure 1.9: Known modularity networks in mastreviruses. Solid lines are protein-protein interactions, dotted lines are DNA-Protein interactions and slashed lines are unknown. Diagram by Darren P Martin (Source Martin et al., 2005, Used with permission)

There are numerous DNA-protein interactions within mastreviruses. Rep, CP, MP, SIR and the LIR all have varying levels of interaction (Figure 1.9). Interestingly the number of interactions each gene has is closely linked to its tolerance to recombination (Figure 1.10; Martin *et al*., 2005). This is important as recombination is the only way to transfer entire modules or sections of modules, and thus is a good indicator of the importance of these interactions to an organism. Tolerance to recombination is also dependent on the

relatedness of a recombinant sequence to the original. This is borne out in mastreviruses, in which MP and SIR are the most modular genes, and they have only two and one interactions respectively. Other segments such as CP have as many as 5 interactions and are far less tolerant to the effects of recombination. It is also notable that only segments of a cassette that are responsible for these interactions will follow these rules. An example of this is the recombination hotspot seen at the CP/SIR interface, which may indicate that the CP genes in this area are less sensitive to intergenome interactions.

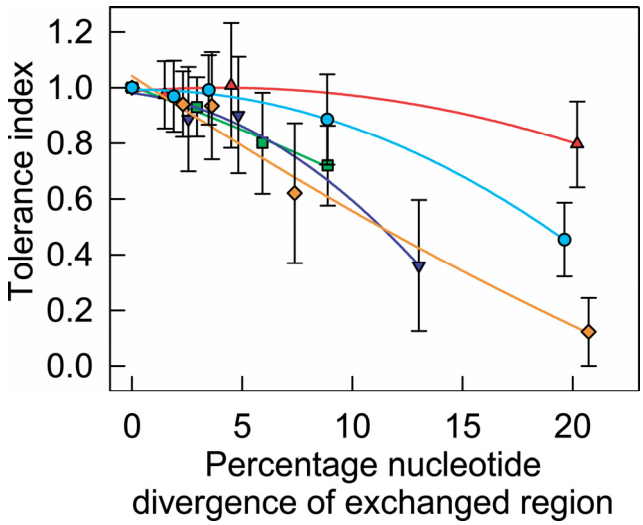


Figure 1.10: Tolerance of the mastrevirus genetic components to recombination. Each plotted point represents a Tolerance index (Ti) value calculated as the average fitness of a pair of recombinant viruses with reciprocally exchanged MP genes (○), CP genes (◆), Rep genes (▼), SIR (▲), or LIR (■) divided by the average fitness of their parental viruses. Error bars represent the standard deviations of Ti values. Curved lines represent quadratic regressions of Ti. (Source Martin et al., 2005, Used with permission)

Perhaps the best understood of these are the CP-MP cassette. There are several known functions of this cassette, and some that are currently only theorised. CP and MP interactions are known to facilitate both inter cell movement of viral DNA, and shuttling of this DNA to the nucleus upon arrival in the host cell (Liu *et al.,* 1997a, 1999, 2001). CP and MP are thought to interact to have some effects on toxicity, and marked decreases in virulence have been observed in viruses with mismatched CP-MP cassettes. Intriguingly transfer of the entire cassette from virulent to non virulent viruses can result in a marked increase in the virulence of the mutant. The exact reasons for the increased toxicity are unknown, as MP is the only inherently toxic protein in mastreviruses.

Currently it is theorized that MP-CP interactions may operate to cause increased expression, thus increasing toxicity by concentration effects.

Modularity then provides another important selection pressure upon mastreviruses. Recombinants that preserve interactions between genome segments will be favoured, and indeed those that enhance virulence will be actively selected for in nature. Conversely disruption of these webs will lead to removal of these recombinants from a population by purifying selection. There is also the possibility of neutral recombinants. Current experiments with recombinants have failed to produce increased fitness over existing natural varieties, but have managed to augment the fitness of less fit strains. The lack of fitness increase would be expected in a natural population that has stabilised over time, and also suggests that negative selection should be the dominant selective force observed in the field.

1.7 Project Aims

This project aims to explore the evolution of mastreviruses, with special interest in the interactions between gene cassettes as evolution occurs. Characterisation of a wide range of mastrevirus isolates is necessary to ensure a meaningful depth of study. The first part of this project aims to characterise a new mastrevirus species and known species in depth. The second part aims to create a database of key conserved sequences and functional areas within the mastreviruses in general. These datasets will then be split into groups based on detectable evolutionary relatedness (i.e MSV-related, SSV-Related)

These groups will then be looked at using bioinformatics tools in order to determine evolutionary patterns within and between groups. A key focus will be examination of the links between various coding regions, in order to determine the extent to which genetic modules are important in mastrevirus evolution. Protein modelling of CP and comparison of conserved areas between species will be used in an attempt to identify functional causes of infectivity, by comparison to field data on virus virulence and infectivity, with the hope of revealing the interactions most responsible for these traits.

2 Molecular Characterisation of a Novel Sugarcane Infecting Mastrevirus Species from South Africa

2.1 Abstract

The sugarcane infecting streak viruses (SISVs) are a diverse collection of mastreviruses (family Geminiviridae) within the African streak virus group. Four SISVs have currently been described, including the well characterized *Maize streak virus* (MSV). Here we present a full annotated sequence record of a new SSIV species, Saccharum streak virus (SacSV) isolated in South Africa. The isolate shares less than 66% identity with any other mastrevirus species, but is most closely related to *Urochloa streak virus* (USV), a species from Nigeria that has until now been an outlier in the African streak virus phylogenetic tree. As with USV, the SacSV isolate we have characterized bares no obvious evidence of inter-species recombination.

This section is based on published material, (**Lawry R, Martin DP, Shepherd DN, van Antwerpen T, Varsani A (2009).** A novel sugarcane-infecting mastrevirus from South Africa. *Arch. Virol.* **154**, 1699-703)

2.2 Introduction

Geminiviruses are a group of plant viruses characterized by their twinned quasi-icosohedral capsid morphology and circular single stranded DNA (ssDNA) genomes. Based on their host ranges, genome organizations and vector specificities geminiviruses are currently classified into four genera: namely begomoviruses, curtoviruses, topocuviruses and mastreviruses. Mastreviruses such as maize streak virus (MSV) and the various sugarcane infecting streak virus (SISV) species found throughout Africa have monopartite genomes, are transmitted by leafhoppers in the Genus *Cicadulina* and infect members of the poacea such as maize and sugarcane. The sugarcane industry in particular is a very important commercial crop, generating both large scale employment and having high value for export, employing over 100,000 people and generating approximately R5 billion in KwaZulu-Natal alone.

The best studied and documented mastreviruses belong to a group known as the African Streak viruses. These viruses infect perennial and annual grass species throught Africa and the South West Indian Ocean islands. African Streak virus species (those with italicised names are currently accepted by the ICTV) include, *Maize streak virus* (MSV); *Panicum streak virus* (PanSV; Varsani *et al*., 2008b), *Sugarcane streak virus* (SSV; Hughes *et al.,* 1993); *Sugarcane streak Egypt virus* (SSEV; Biggare *et al.,*1999*)*, *Sugarcane streak Reunion virus* (SSRV; Shepherd *et al.,* 2008), Eragrostis streak virus (ESV; Shepherd *et al.,* 2008) and. Urochloa streak virus (USV; Oluwafemi *et al.,* 2008*)*.

Four of these seven species, (MSV, SSV, SSEV, SSRV) are known to produce disease in sugarcane that is potentially of economic significant (Biggare *et al.,*1999, Hughes *et al.,* 1993, van Antwerpen *et al.,* 2008). When one considers both that these four sugarcane infecting species are highly divergent and that economically relevant disease in maize is caused by only one of the 5 known strains of the virus and that the virulent strain is genetically homogenous (Varsani *et al*.,2008a), it is apparent that sugarcane (or at least some genotypes of this species) is perhaps particularly sensitive to being infected with African streak viruses. Both maize and sugarcane are species that were only introduced

to Africa during its colonization by Europeans over the last five centuries and it has therefore long been assumed that viruses that have emerged in these species have done so from indigenous African wild-grass species. In support of this MSV, SSRV and SSV strains have been widely found in wild-grass species (Shepherd *et al.,* 2008). Whereas the MSVs have generally been found in *Digitaria*, *Setaria* and *Urochloa* species (Varsani *et al*., 2008a) SSV and SSRV have been found in *Setaria barbata* (East Indian bristlegrass)*, Cenchrus myosuroides* (big sandbur), *Paspalum conjugatum* (hilograss) and *Eragrostis curvula* (weeping lovegrass; Shepherd *et al.,* 2008).

2.3 Methods

2.3.1 Sequencing and Isolation

The new species was discovered during routine screening of sugarcane fields for evidence of MSV infections in the Kwa-Zulu-Natal province of South Africa (Lat -28.737285; .Lon – 31.885717). Total DNA was extracted from a sugarcane cultivar (N44) presenting with streak symptoms similar to those produced by MSV using a CTAB-based protocol developed by Doyle & Doyle (Doyle & Doyle., 1987). 7 ml of CTAB isolation buffer (2% hexadecyltrimethylammoniumbromide, 1.4 M NaCl, 0.2% 2-mercaptoethanol, 20 mM EDTA, 100mM Tris-HCL, pH 8.0) was preheated to 60 degrees. One gram of ground leaf tissue was added and incubated at 60 degrees for 30 minutes. The suspension was then extracted using a 24:1 chloroform-isoamyl alcohol solution. This was then centrifuged at 6000 x g for 10 minutes. Nucleic acids were pelleted using isopropanol and centrifuged at low g for 2 minutes. This was then resuspended in 1ml TE (10 mM Tris-HCL, 1 mM EDTA, pH 7.4) RNase A was added to 10 υμg/ml and incubated for 30 minutes at 37°C degrees Celsius. DNA was then precipitated using Ethanol, and spun down at 10000 x g for 10 minutes and resuspended in TE.

A full length mastrevirus genome was isolated/amplified using ϕ29 DNA polymerase amplification (TempliPhi™, GE Healthcare) (Owor *et al.,* 2007a, 2007b). 2.5 μl of the

DNA solution was added to 2.5 µl of the sample buffer from the kit. This was heated to 95 degrees Celsius for 3 minutes. This was then cooled to room temperature and added to 5 µl of reaction buffer and 0.2 µl of enzyme mix from the kit. This was then incubated for 20 hours at 30 degrees Celsius. The resulting concatamers were digested with *Bam*HI to yield full length monomeric (~2.7 Kb) full-genome length molecules. *Bam*HI was chosen as it is known to digest MSV genomes at approximately full length. These full length molecules were then ligated to *Bam*HI digested pGEM3 Zf(+) (Promega Biotech) and sequenced at Macrogen Inc. (Korea) using primer walking.

2.3.2 Sequence Alignment

The assembled genome (2,744 bp) sequence was aligned with those of well characterized mastreviruses, including MSV strains A to K (Varsani *et al*., 2008a), PanSV strains A to D (Varsani *et al.,* 2008b), USV (Oluwafemi *et al.,* 2008*)*, ESV (Shepherd *et al.,* 2008) SSRV, SSEV and SSV (Shepherd *et al.,* 2008). The sequences were aligned using Clustal W (gap open penalty = 10; gap extension penalty = 5) (Thompson *et al.,* 2004*)*, followed by manual alignment in Mega 4.0 (Tamura *et al.,* 2007). Open Reading frames were identified by homology to known sequences available on the BLAST database. Once identified sequence identity to the representative African streak virus sequences of the full genome and various open reading frames were calculated using MEGA 4 (Tamura *et al.,* 2007) with pairwise deletion of gaps.

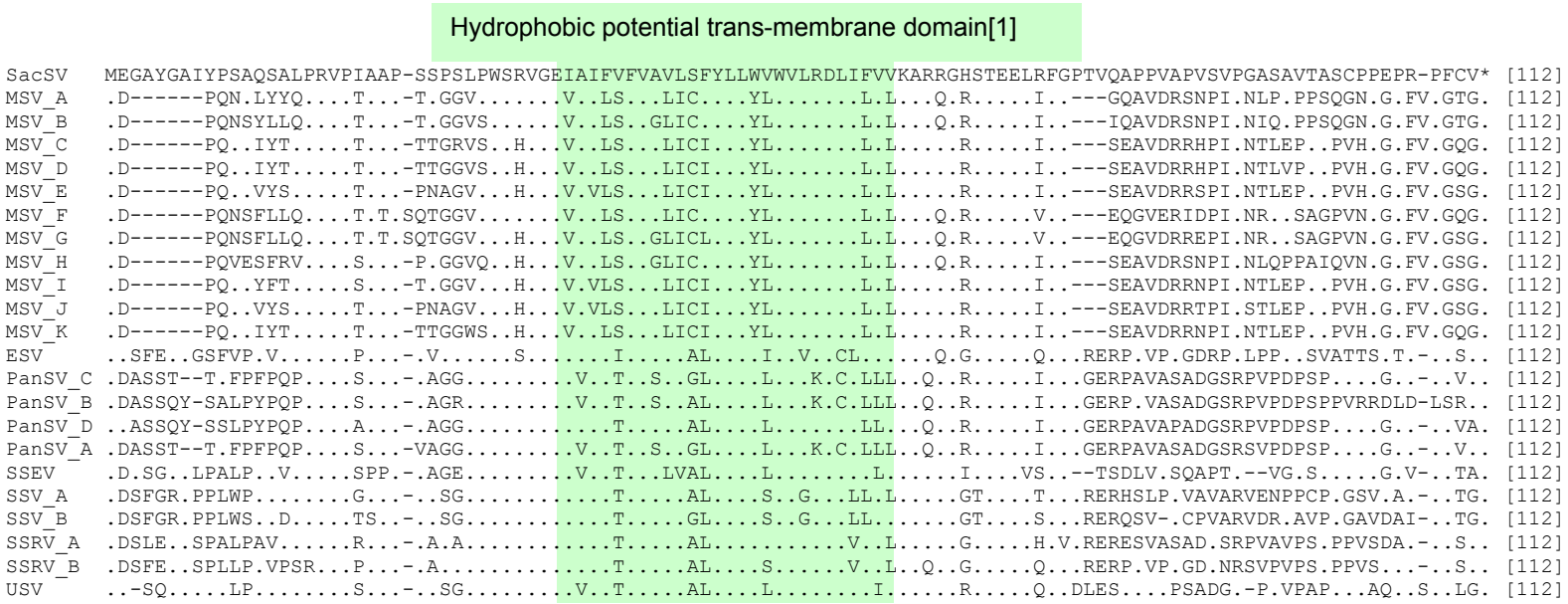
2.3.3 Recombination and Phylogenetics

Due to the extensive recombination observed in single stranded DNA viruses (Lefeurve *et al,.* 2009), we analysed the new isolate along with a diverse array of MSV, PanSV, ESV, USV, SSEV, SSV and SSRV strains (see Table 1 for details) for evidence of recombination using RDP, GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN and 3SEQ implemented in RDP3 (Martin., (2009). Maximum likelihood phylogenetic trees (Figure 2.15) were drawn using PHYML (Guindon *et al,.*2003*)* using the model, GTR+I+G4, automatically selected as the best-fit model using RDP3 (Lefeurve *et al,.* 2009).

2.4 Results

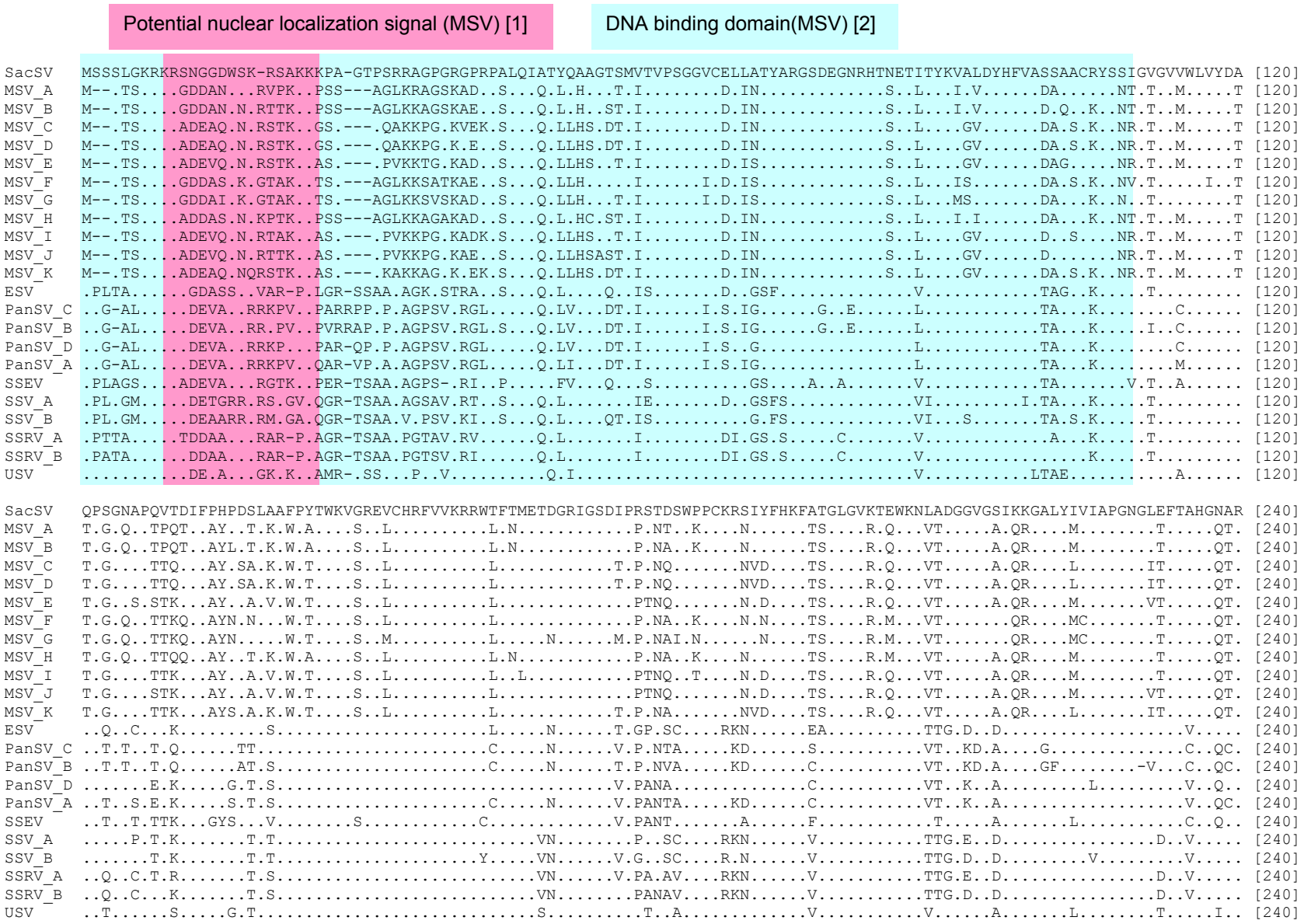
2.4.1 Genetic Structure

We identified probable movement protein (*MP*), coat protein (*CP*) and replication associated protein genes (*Rep* and *RepA*) (*Fig 2.1-2.3)*. We also identified the canonical virion strand origin of replication (v-*ori*) TAATATTAC sequence and its associated conserved inverted repeat sequences. TATA boxes for virion and complementary sense transcription and GC rich sequences (close to the v-*ori*; similar to the MSV rightward promoter elements) (Fenoll *et al.,* 1987,1990) were also identified (*Fig2. 4*). We identified the probable intron and spliced replication protein product (*Fig 2.3*). The predicted amino acid sequences of expressed CP, MP and Rep proteins are provided in Figures 2.1-4. The Rep of SacSV contains the **L**H**C**Y**E** (**L**x**C**x**E**) binding motif at residues 204-208 and the rolling circle motifs **FLTYS** (residue 25-29) and **YILK** (residues 107-110; *Fig 2.3*). The potential NTP binding motif (**VG**X_4**GKTS**X_{29}**DD**) in the SacSV Rep is located between residues 233-275.



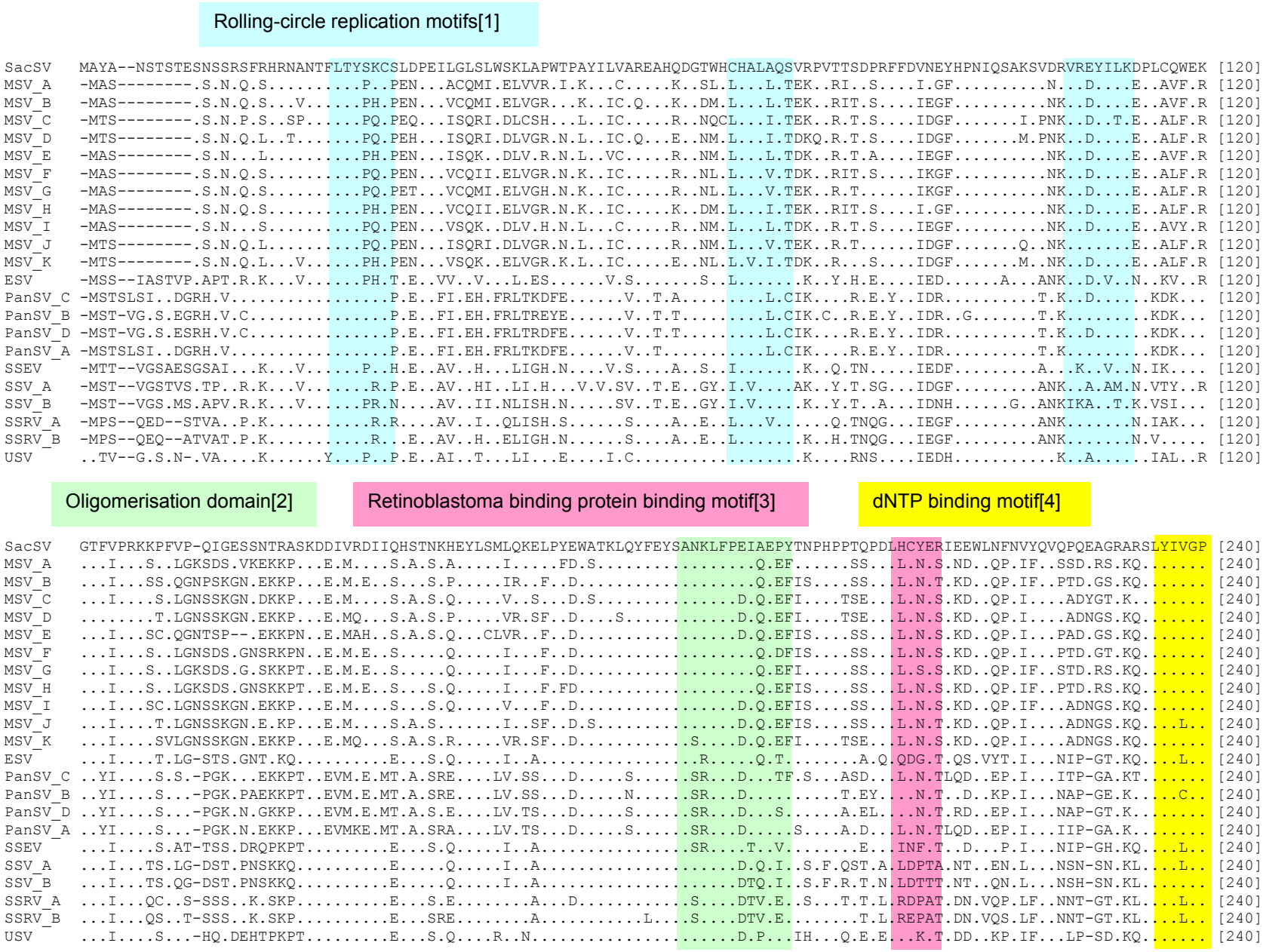

```
SacSV     MEGAYGAIYPSAQSALPRVPIAAP-SSPSLPWSRVGEIAIFVFVAVLSFYLLWVWVLRDLIFVVKARRGHSTEELRFGPTVQAPPVAPVSVPGASAVTASCPPEPR-PFCV* [112]
MSV_A     .D------PQN.LYYQ....T...-T.GGV.......V..LS...LIC....YL.......L.L...Q.R.....I..---GQAVDRSNPI.NLP.PPSQGN.G.FV.GTG. [112]
MSV_B     .D------PQNSYLLQ....T...-T.GGVS......V..LS..GLIC....YL.......L.L...Q.R.....I..---IQAVDRSNPI.NIQ.PPSQGN.G.FV.GTG. [112]
MSV_C     .D------PQ..IYT.....T...-TTGRVS..H...V..LS...LICI...YL.......L.L.....R.....I..---SEAVDRRHPI.NTLEP..PVH.G.FV.GQG. [112]
MSV_D     .D------PQ..IYT.....T...-TTGGVS..H...V..LS...LICI...YL.......L.L.....R.....I..---SEAVDRRHPI.NTLVP..PVH.G.FV.GQG. [112]
MSV_E     .D------PQ..VYS.....T...-PNAGV...H...V.VLS...LICI...YL.......L.L.....R.....I..---SEAVDRRSPI.NTLEP..PVH.G.FV.GSG. [112]
MSV_F     .D------PQNSFLLQ....T.T.SQTGGV.......V..LS...LIC....YL.......L.L...Q.R.....V..---EQGVERIDPI.NR..SAGPVN.G.FV.GQG. [112]
MSV_G     .D------PQNSFLLQ....T.T.SQTGGV...H...V..LS..GLICL...YL.......L.L...Q.R.....V..---EQGVDRREPI.NR..SAGPVN.G.FV.GSG. [112]
MSV_H     .D------PQVESFRV....S...-P.GGVQ..H...V..LS..GLIC....YL.......L.L...Q.R.....I..---SEAVDRSNPI.NLQPPAIQVN.G.FV.GSG. [112]
MSV_I     .D------PQ..YFT.....S...-T.GGV...H...V.VLS...LICI...YL.......L.L.....R.....I..---SEAVDRRNPI.NTLEP..PVH.G.FV.GSG. [112]
MSV_J     .D------PQ..VYS.....T...-PNAGV...H...V.VLS...LICI...YL.......L.L.....R.....I..---SEAVDRRTPI.STLEP..PVH.G.FV.GSG. [112]
MSV_K     .D------PQ..IYT.....T...-TTGGWS..H...V..LS...LICI...YL.......L.L.....R.....I..---SEAVDRRNPI.NTLEP..PVH.G.FV.GQG. [112]
ESV       ..SFE..GSFVP.V......P...-.V......S.......I.....AL....I..V..CL.......Q.G.....Q...RERP.VP.GDRP.LPP..SVATTS.T.-..S.. [112]
PanSV_C   .DASST--T.FPFPQP....S...-.AGG.........V..T..S..GL....L...K.C.LLL..Q..R.....I...GERPAVASADGSRPVPDPSP....G..-..V.. [112]
PanSV_B   .DASSQY-SALPYPQP....S...-.AGR.........V..T..S..AL....L...K.C.LLL..Q..R.....I...GERP.VASADGSRPVPDPSPPVRRDLD-LSR.. [112]
PanSV_D   ..ASSQY-SSLPYPQP....A...-.AGG.............T.....AL....L.......LL...Q..R.....I...GERPAVAPADGSRPVPDPSP....G..-..VA. [112]
PanSV_A   .DASST--T.FPFPQP....S...-VAGG.........V..T..S..GL....L...K.C.LLL..Q..R.....I...GERPAVASADGSRSVPDPSP....G..-..V.. [112]
SSEV      .D.SG..LPALP..V.....SPP.-.AGE.........V..T...LVAL....L........L......I....VS..--TSDLV.SQAPT.--VG.S.....G.V-..TA. [112]
SSV_A     .DSFGR.PPLWP........G...-..SG.............T.....AL....S..G...LL.L.....GT....T...RERHSLP.VAVARVENPPCP.GSV.A.-..TG. [112]
SSV_B     .DSFGR.PPLWS..D.....TS..-..SG.............T.....GL....S..G...LL.......GT....S...RERQSV-.CPVARVDR.AVP.GAVDAI-..TG. [112]
SSRV_A    .DSLE..SPALPAV......R...-.A.A.............T.....AL...........V..L.....G.....H.V.RERESVASAD.SRPVAVPS.PPVSDA.-..S.. [112]
SSRV_B    .DSFE..SPLLP.VPSR...P...-.A...............T.....AL....S......V..L..Q..G.....Q...RERP.VP.GD.NRSVPVPS.PPVS...-..S.. [112]
USV       ..-SQ.....LP........S...-..SG.........V..T.....AL....L........I......R.....Q..DLES....PSADG.-P.VPAP...AQ..S..LG. [112]
```

Figure 2.1: Annotated predicted movement protein amino acid sequences of SacSV [ZA-Emp-T1-2008] together with a selection of major strain variants from other publicly available African streak virus species (GenBank accession numbers are provided in Table 2.1). The hydrophobic, potentially membrane-spanning internal domain of the sequences is highlighted. Wherever amino acids in a particular alignment column are identical to that of SacSV [ZA-Emp-T1-2008], they are replaced with a "." character. Gaps introduced to optimise the alignment are indicated with a "-" character. The character "*" represents a stop codon.Protein annotation of the movement proteins of SacSV and a representative sample of mastreviruses, Note the potential trans-membrane domain which has a high number of hydrophobic amino acids.


Potential nuclear localization signal (MSV) [1]
DNA binding domain(MSV) [2]

```
SacSV   LYFKSVGNQ* [250]
MSV_A   .......... [250]
MSV_B   .......... [250]
MSV_C   .......... [250]
MSV_D   .......... [250]
MSV_E   .......... [250]
MSV_F   .......... [250]
MSV_G   .......... [250]
MSV_H   .......... [250]
MSV_I   .......... [250]
MSV_J   .......... [250]
MSV_K   .......... [250]
ESV     .......... [250]
PanSV_C .......... [250]
PanSV_B .......... [250]
PanSV_D .......... [250]
PanSV_A .......... [250]
SSEV    .......... [250]
SSV_A   .......... [250]
SSV_B   .......... [250]
SSRV_A  .......... [250]
SSRV_B  .......... [250]
USV     .......... [250]
```

Figure 2.2: Annotated predicted coat protein amino acid sequences of SacSV [ZA-Emp-T1-2008] together with a selection of major strain variants from other publicly available African streak virus species (GenBank accession numbers are provided in Table 2.1). The potential nuclear localization signal, and DNA binding domains (inferred by analogy with those determined for MSV) are highlighted on the sequence. Wherever amino acids in a particular alignment column are identical to that of SacSV [ZA-Emp-T1-2008], they are replaced with a "." character. Gaps introduced to optimise the alignment are indicated with a "-" character. The character "*" represents a stop codon.Protein annotation of the coat protein of SacSV and a representative sample of mastreviruses. The DNA binding domain and nuclear localization signal allow for non specific binding of both ss and dsDNA. Also note the significant number of basic amino acids, which may potentially enhance DNA binding.


Rolling-circle replication motifs[1]
Oligomerisation domain[2]
Retinoblastoma binding protein binding motif[3]
dNTP binding motif[4]

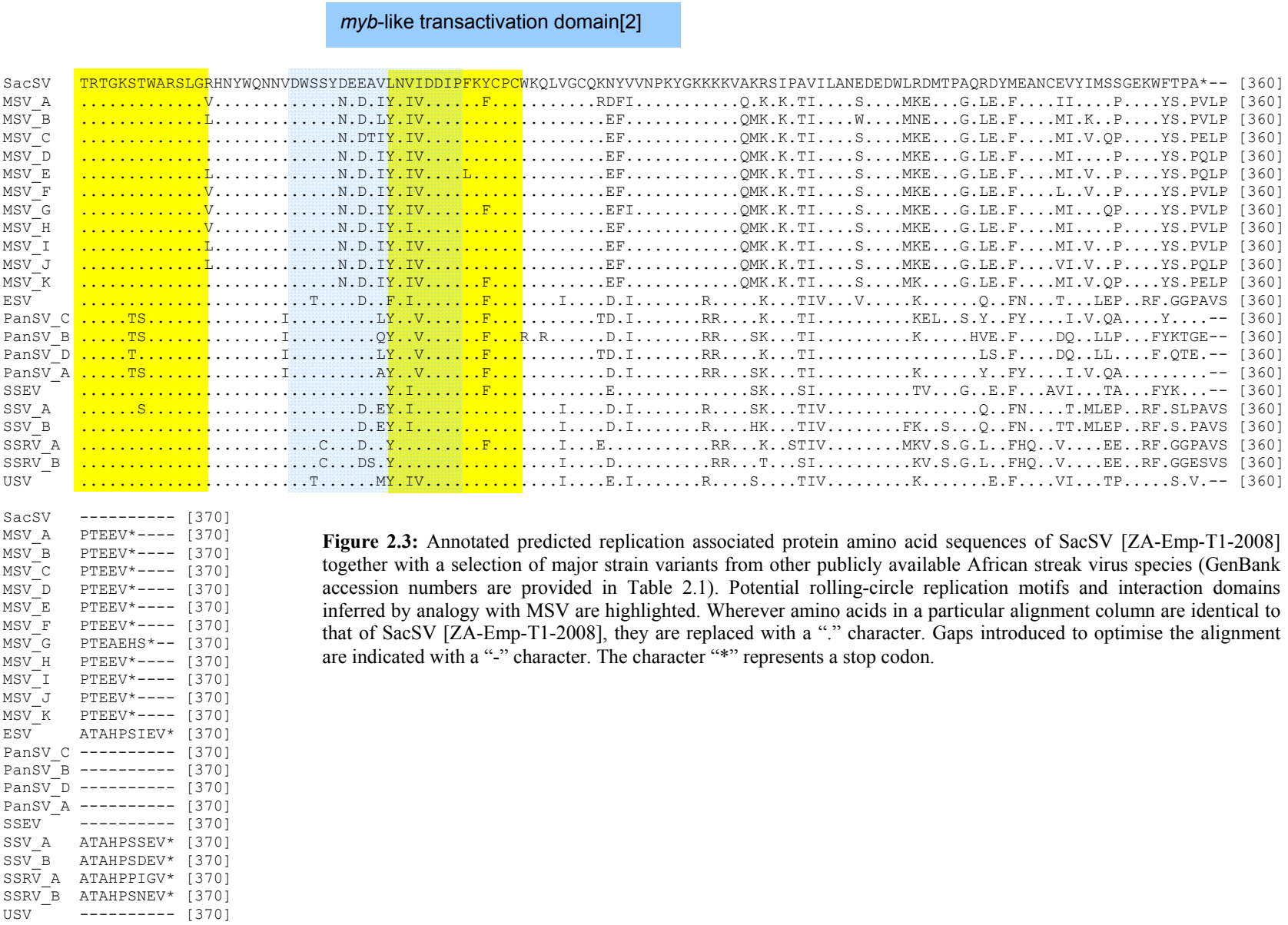
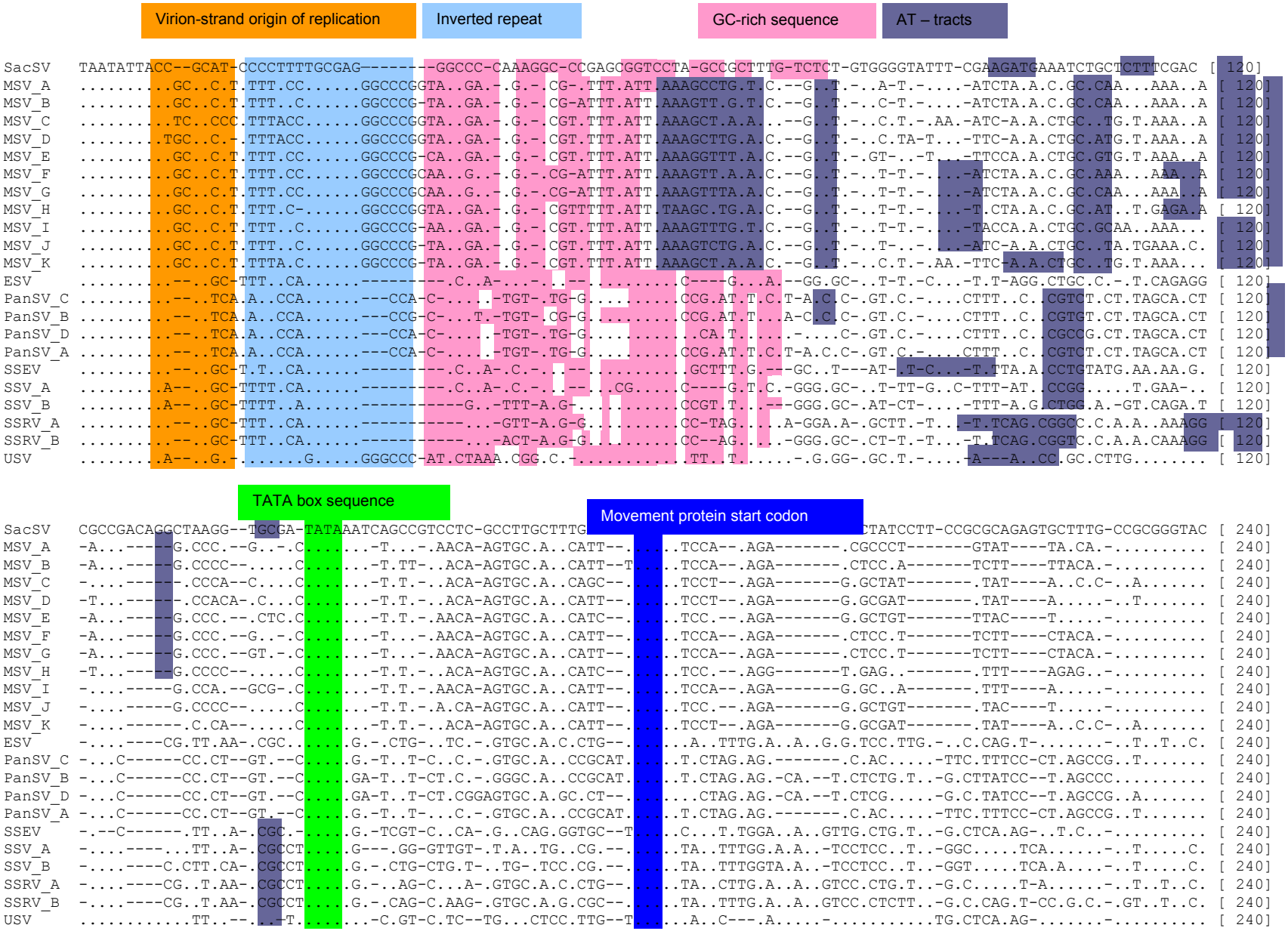
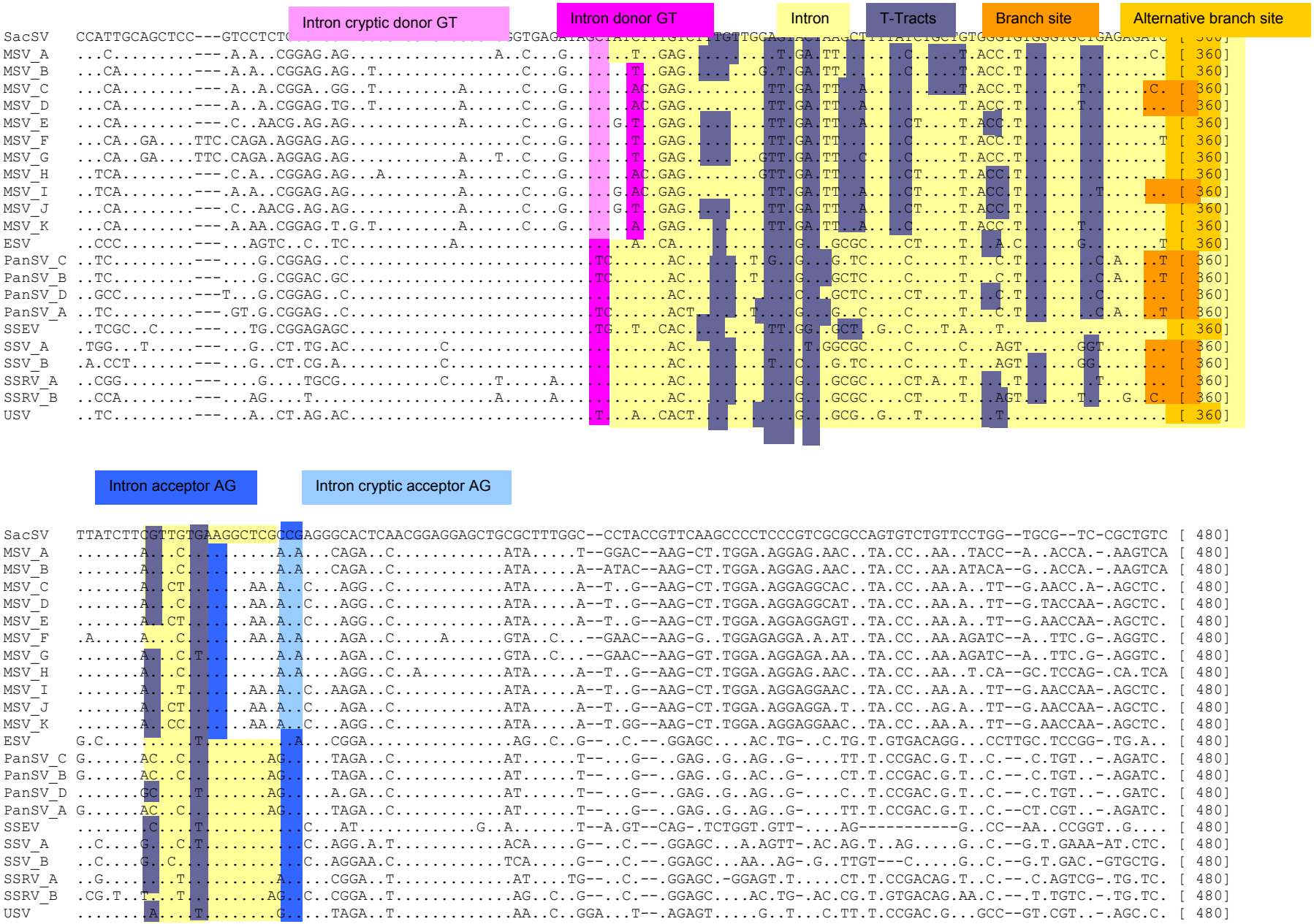
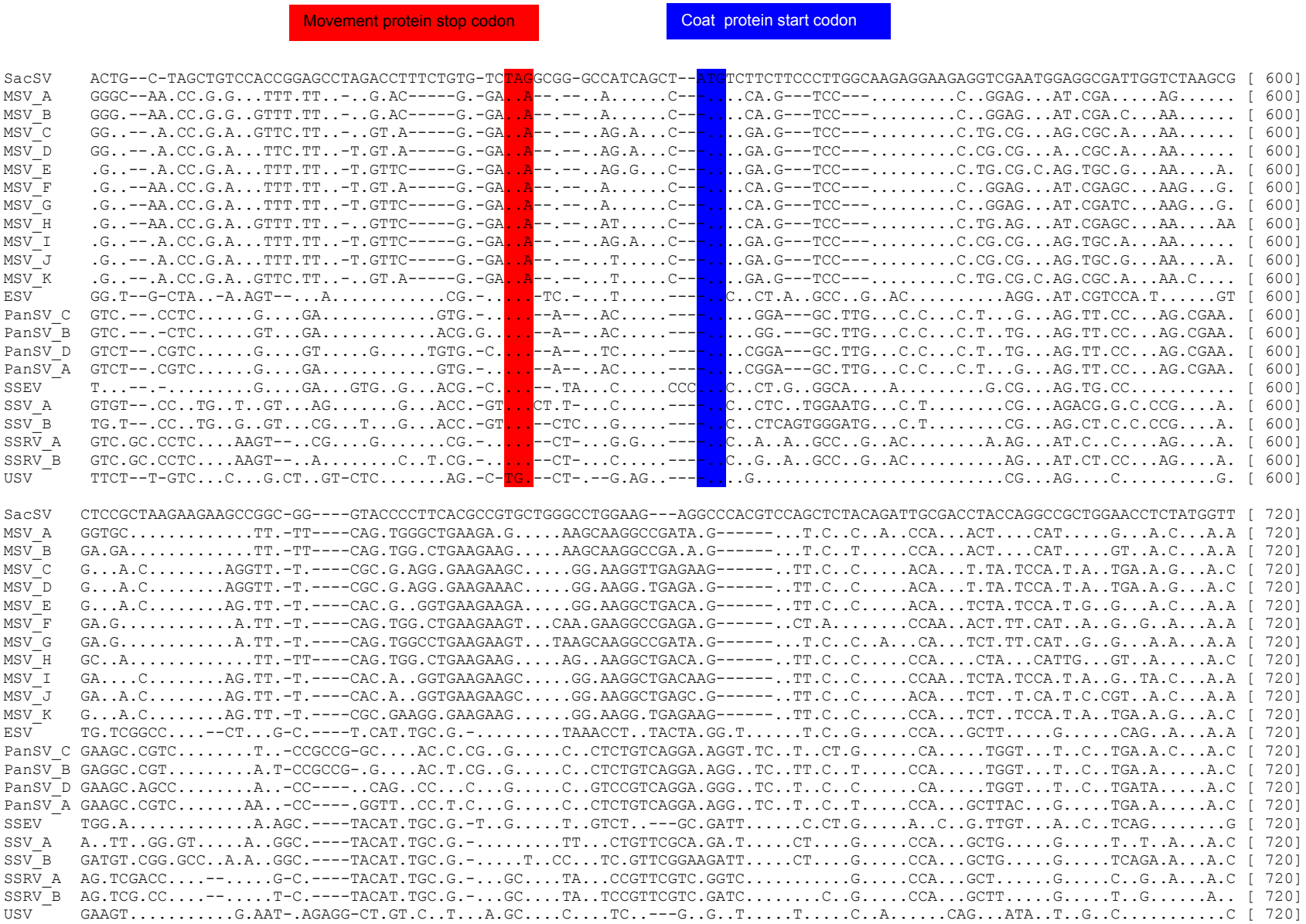


Figure 2.3: Annotated predicted replication associated protein amino acid sequences of SacSV [ZA-Emp-T1-2008] together with a selection of major strain variants from other publicly available African streak virus species (GenBank accession numbers are provided in Table 2.1). Potential rolling-circle replication motifs and interaction domains inferred by analogy with MSV are highlighted. Wherever amino acids in a particular alignment column are identical to that of SacSV [ZA-Emp-T1-2008], they are replaced with a "." character. Gaps introduced to optimise the alignment are indicated with a "-" character. The character "*" represents a stop codon.

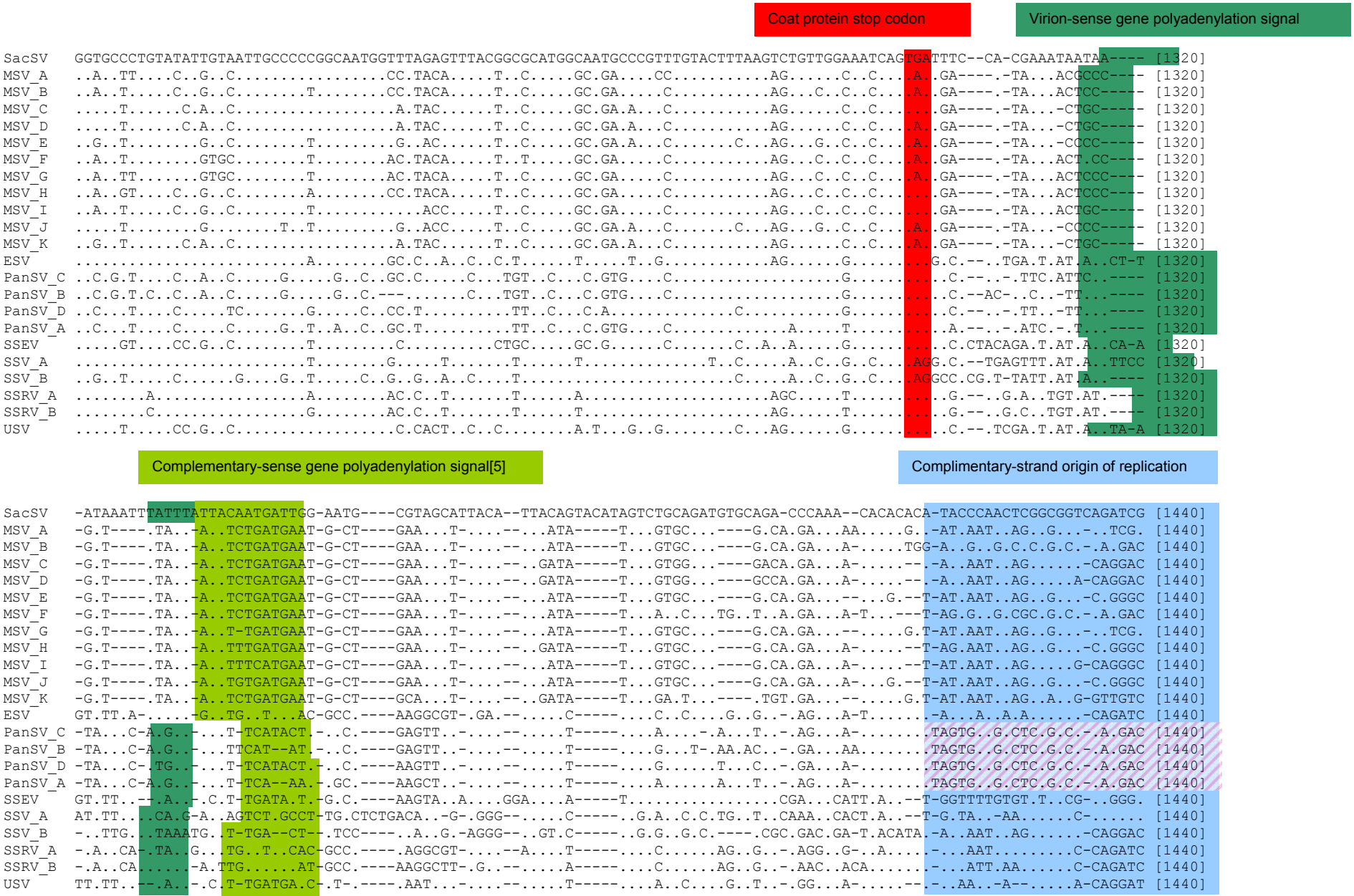

Virion-strand origin of replication
Inverted repeat
GC-rich sequence
AT – tracts
TATA box sequence
Movement protein start codon

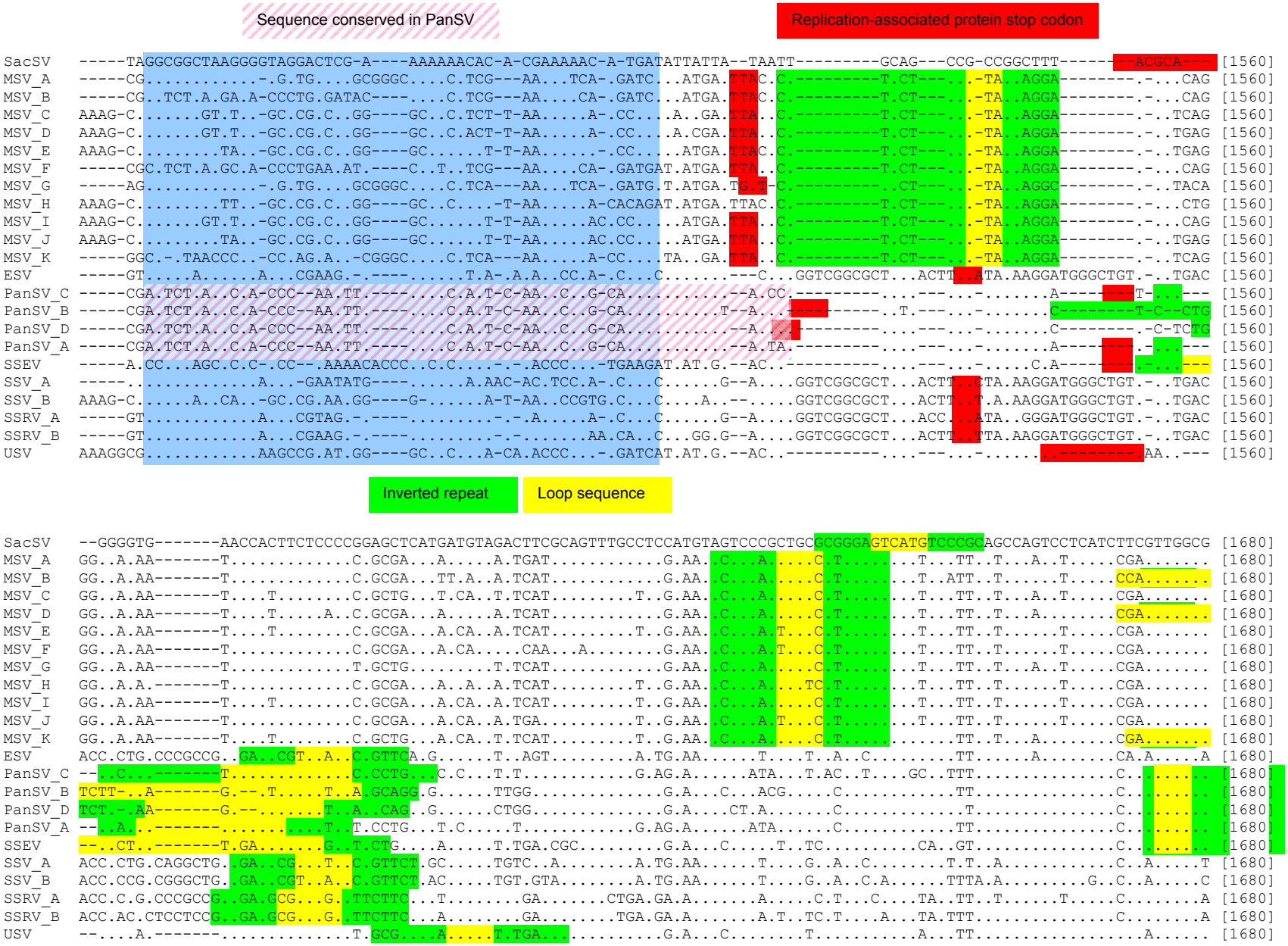

Intron cryptic donor GT
Intron donor GT
Intron
T-Tracts
Branch site
Alternative branch site
Intron acceptor AG
Intron cryptic acceptor AG

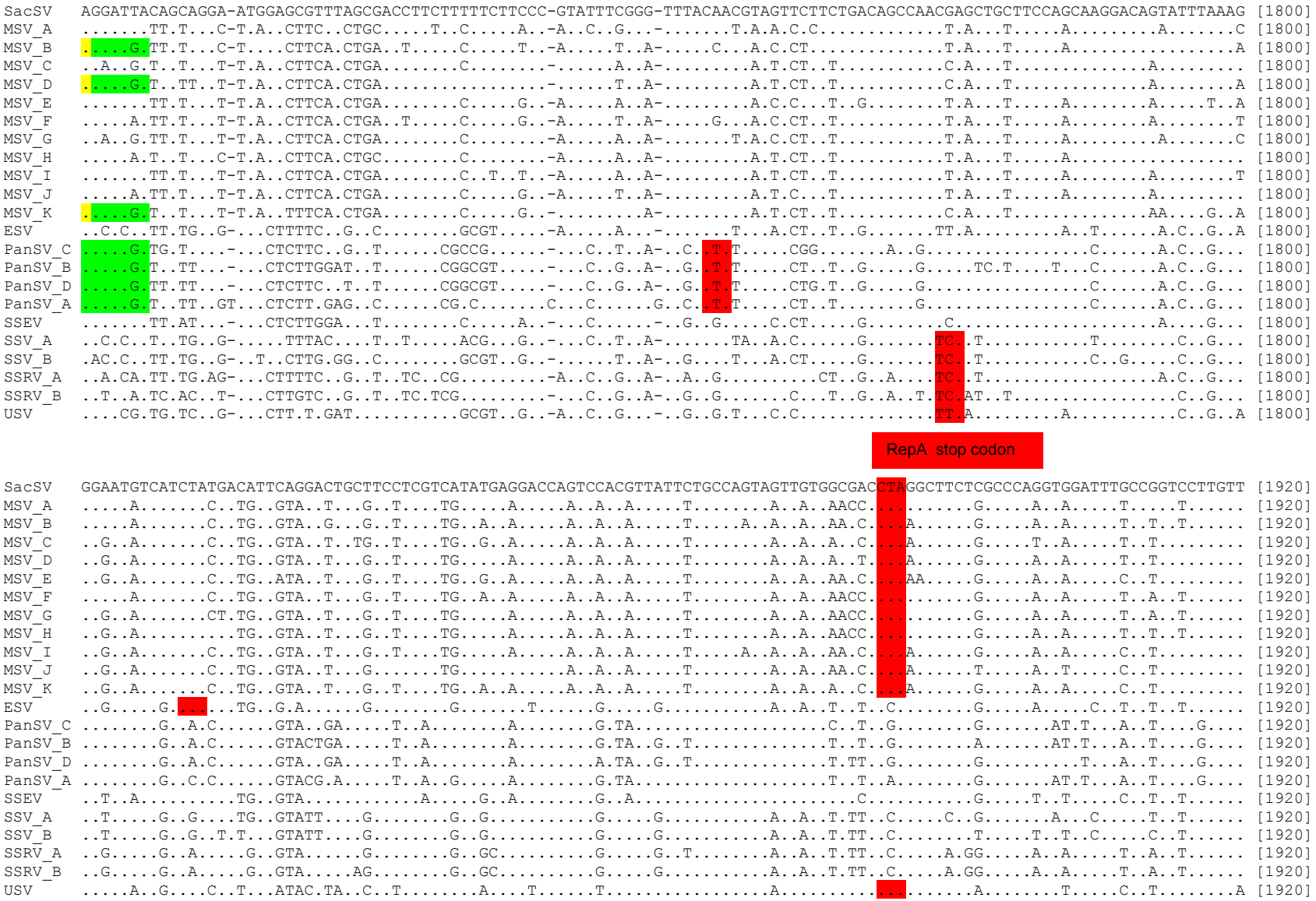

Movement protein stop codon
Coat protein start codon

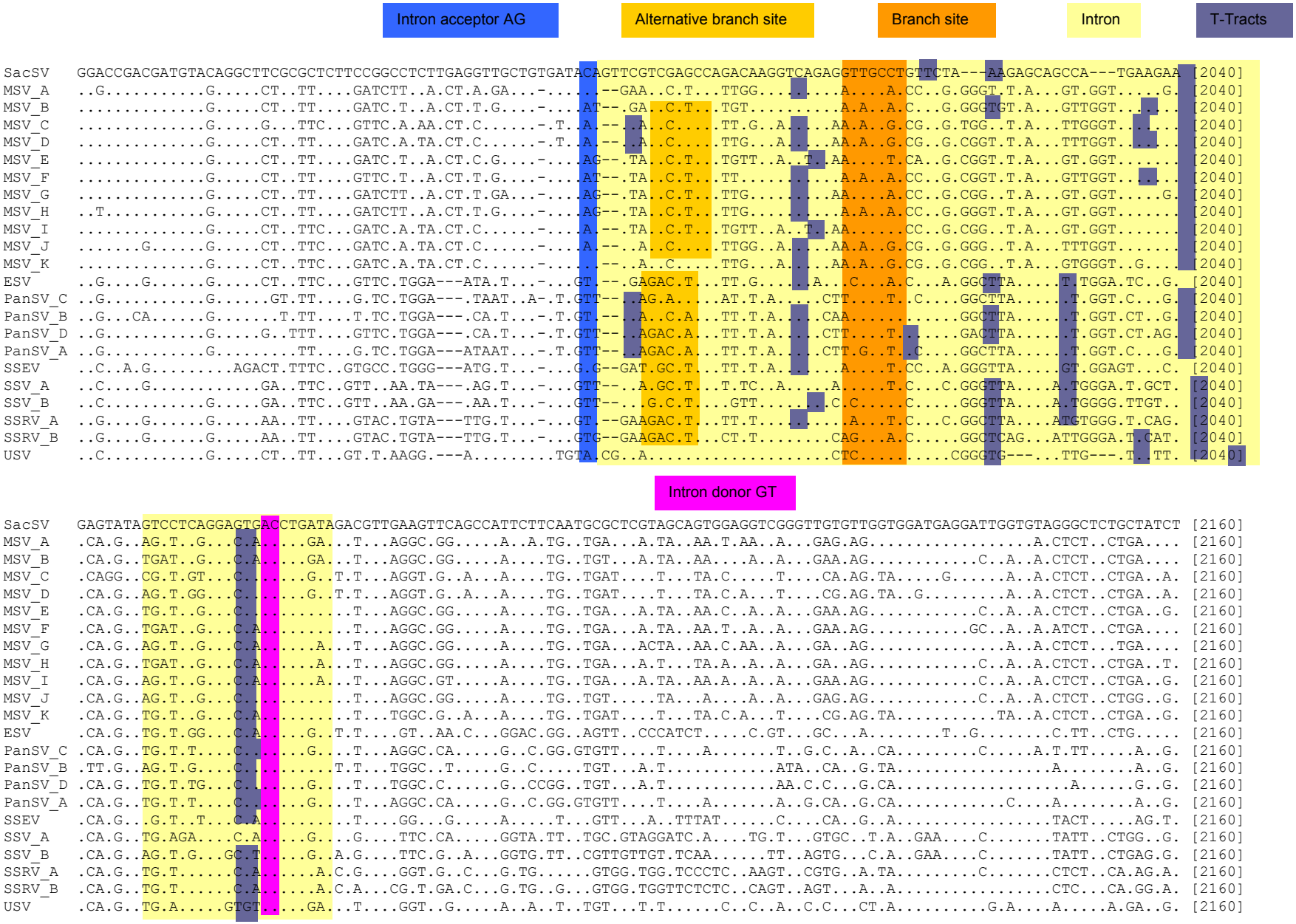
```
SacSV   ACTGTCCCTAGCGGGGGCGTTTGTGAACTCCTTGCGACATATGCTCGAGGGTCCGACGAGGGCAACCGTCACACCAACGAGACTATCACGTACAAGGTTGCCTTGGACTACCACTTTGTA [ 840]
MSV_A   ........ATC...A..A..A.....C...A.CAAC..C.....C.....A..T.............C.......G.......C.G.........A.C...G.C...........C..T [ 840]
MSV_B   ..A.....CTC...A..A..A.....C...A.CAAC..C.....C.....A.................C.......G.......C.G.........A.....G.C...........C..T [ 840]
MSV_C   ..C..A...TCA..T..G..C.....C...A.CAAT........C..........T.................G...A..CC.A.........A....GGG.C..T........C..T [ 840]
MSV_D   ..C..A..GTCA..T..G..C.....C...A.CAAT........C........T..T.............G.......G...A..CC.A.........A....GGG.C...........C..C [ 840]
MSV_E   ..C..A..GTCA..T..G..C.....C...A.CAGT........C.....A.....T.................G...A..CC.A.........A....GGG.C...........C..C [ 840]
MSV_F   ..A.....CTC...A..AA....C..C..AA.CAGC..C.....C.....T..A..T........T........G.G.......C.T.........A.AT..C.C...........C... [ 840]
MSV_G   ..A.....CTC...A..TA....C..C...A.CAGC..C.....C.....T..A..T...........A.......G.......C.T.........A.GT..C.C...........C..T [ 840]
MSV_H   .....G..ATCT.....T..C.....C...A.CAAT..C.....C.....A.....T...........C.......G.......C.G.........A.C...A.C...........C..T [ 840]
MSV_I   ..C..A..GTCA..T..G..C.....C...A.CAAC........C.....T..A..T........T..A.......G...A..CC.A.........A....GGG.C...........C..T [ 840]
MSV_J   ..C..A..GTCA..T..G..C.....C...A.CAAC........C..........T...........G.......G...A..CC.A.........A....GGG.C..T........C..C [ 840]
MSV_K   ..C..A..GTC...T..G..C.....C...A.CAAT........C..........T...........G.......G...A..CC.A.........A....GGG.C...........C..C [ 840]
ESV     T.......GTCT..T...........TT.G....GCT.G.T...C.....A.....T...........C..............G.A..A........A...C.C.....T..T..C..G [ 840]
PanSV_C .....G..GTCA..C...A.C..CAGT..GA...GC..G.....C..G..C..A.GT..A..TG.A..C..............C.G..............TC............C..T [ 840]
PanSV_B .....G..GTC...A...A.C..CAGT..GA...GC..G.....C..G..C..T.GT..A..TG.A..C..............C.G...........C..TC............C..T [ 840]
PanSV_D .....G..GTCA..C...A.C..CAGT..G....GC..C.....C..C..A..T..T..A..A.....................C.A.............G..A............T [ 840]
PanSV_A ..G..G..GTC...C...A.C..CAGT..GA...GC..G.....C..G..C..T..T..A..T.....C..............C.G..............GC....T........C..T [ 840]
SSEV    T........TC...T..G..........G..G.GCT.C...........CG.T.......C.....................G...............................C..T [ 840]
SSV_A   GAA..G..ATC...T...........C..T....GTT.C.TCT.C..T..C..T........G...A.G..............G...TC........G..AC............CA.T [ 840]
SSV_B   T.......GTC.................T....GT..C.TCT.C..T..C..T........G...A.G..............G...TC..T.....GT.AC............C..T [ 840]
SSRV_A  ..C..T..CTCG..T..T........TA....C.G.T.C...T.A.....C..T......T....T..C..............G....C........A..TC..........T..C..T [ 840]
SSRV_B  ..C.....CTCG..............TA.T....GTT.C..CT.A.....C..T......T.....................G.T..C........G..TC.C........T.....T [ 840]
USV     .....T..GTC...T...........G...........C..C..AA....A.....T.....T...A.G..............G.T...........C...C....T........C..G [ 840]

SacSV   GCCTCGTCTGCTGCTTGCAGGTACTCTTCCATCGGTGTGGGGGTCGTGTGGTTGGTGTATGATGCACAGCCCTCCGGCAATGCCCCCCAGGTAACGGATATCTTCCCGCATCCTGATAGT [ 960]
MSV_A   ...GACG.G.....C...C.C.....CAA..C...AACC..T..AA.....C..........CA.CACT...GG...AC.A..T..GACCCCGCAAAC...A..TG.CT.C.....C.CG [ 960]
MSV_B   ..AGAC..GCAA..C....A...T...AA..C...GACC..T..GA.....C.......C..CA.CACT...GG...AC.A..G..GACCCC.CAAACG..A..TG.AT...TG..C.CG [ 960]
MSV_C   ..TGACG.C..AT.C....A......CAA.CG...AACA..T..GA.....C.......C..CA.GACT...GG...A..C..A...ACCACCCAA.....T...G.AT.C..CTCCGCA [ 960]
MSV_D   ..TGACG.C..AT.C....A......CAATCG...AACA..T..GA.....C.......C..CA.TACT...GG...A..C..A...ACCACCCA......T...G.AT.C..CTCCGCA [ 960]
MSV_E   ...GACG.A.GA..C.....A.....CAA.CG...AACA..T..GA.....C.......C..CA.TACC...GG...A..CT.A..GTCCACG.A...C......G.CT....G...GCA [ 960]
MSV_F   ...GACG.A..AT.C....A......CAA.G.T..GACT..C..G......C.CA.C..C..CA.GAC....GG...AC.A..G..GAC.AC..AAC.G..A..TG.CT.CAAC..C.AC [ 960]
MSV_G   ..TGACG.A..A..G....A......CAA...T..AACC..C..G..C...C....T..C..CA.TACT...GG...AC.A.....GACTACC.AAC.G..A..TG.CT.CAAC..C..C [ 960]
MSV_H   ..TGACG.C..A..C....A...T..CAA..C...GACC..T..GA.....C..........CA.TACT...GG...AC.A..G..GACCACGCAAC.G..A..TG.CT.C..G..C.CG [ 960]
MSV_I   ..AGAC......T.......A.....CAA.CG...AACA..T..GA.....C....T..C..CA.TACT...GG...A.....T..AACTACG.A...C......G..T....G..CGCG [ 960]
MSV_J   ..AGAC.....A..A...........CAA.CG....ACA..T..GA.....C.......C..CA.TACC...GG...A.....A..ATCCACG.A...C......G.CT....G..CGCA [ 960]
MSV_K   ..TGACG.C..AT......A......CAA.CG...AACA..T..GA.....C.......C..CA.TACT...GG...A..C..A...ACCACC.AA.........G.AT.CT.G..CGCA [ 960]
ESV     ..TA.TG.A.G...C....A...T..AAGT.....CACT..C..G..................C........TCAG..G..CTGT..........A......T..............TCC [ 960]
PanSV_C ..TA.CG.C..G..C....A......CAG...T..A..C..T..GTGC...........C.....G.....TA......C......GACT..GCA...C.....T.....C..GAC..CG [ 960]
PanSV_B ...A.CG.......G..T.AA...AGCAGT..T..CA.T..A..GTGC...........C.....G.....TA......C......GACC..CCAA..C..T..T..T..C..G.C..C. [ 960]
PanSV_D ...A.CG.......G..T.AA...AGCAGT..T.....T..A..GTGC...........C.....G.....T..............GG....G.A............T...GGA....CG [ 960]
PanSV_A ..TA.TG....G..C....A......CAG...T..A.....T...A...................G.....GA........T.A..GG....G.A...-.........T..CT.C..C.CG [ 960]
SSEV    ..TA.TG.A..C..C........T..CAG.G.T..CACT..C..G.CC.....................TA....A...A....GAC.ACG.A...C......GGCT..AG...CTCC [ 960]
SSV_A   ..TA.CG.......C....A...T..A..T..T..CACT..C..G.....................C........T..G..G...C....GACA..G.A......T.....T..C.......C. [ 960]
SSV_B   ...A.CG.C...T.C....A...T..A..G.....CACT..C..G..............C..C........T..G..G........GACA..G.A.........T..T..C..G....C. [ 960]
SSRV_A  ..TAGTG.A..A..C....A...T..CAGT..T..CACT..C..G..............C..C........TCAG......TGT..TAC...G.GA..C..................C.C. [ 960]
SSRV_B  ..GAGT..A..A..C....A......GAGT..T..CACT..C..G.........................TCAG..T...TG...G.....T.A...C........A.....A..C.C. [ 960]
USV     CT.A.AG.A.AG..C...C.C.....G...........A.....G.CT..............C.........A.......C.....T.....CT.T..C..A.....A..CGGC..C.C. [ 960]
```

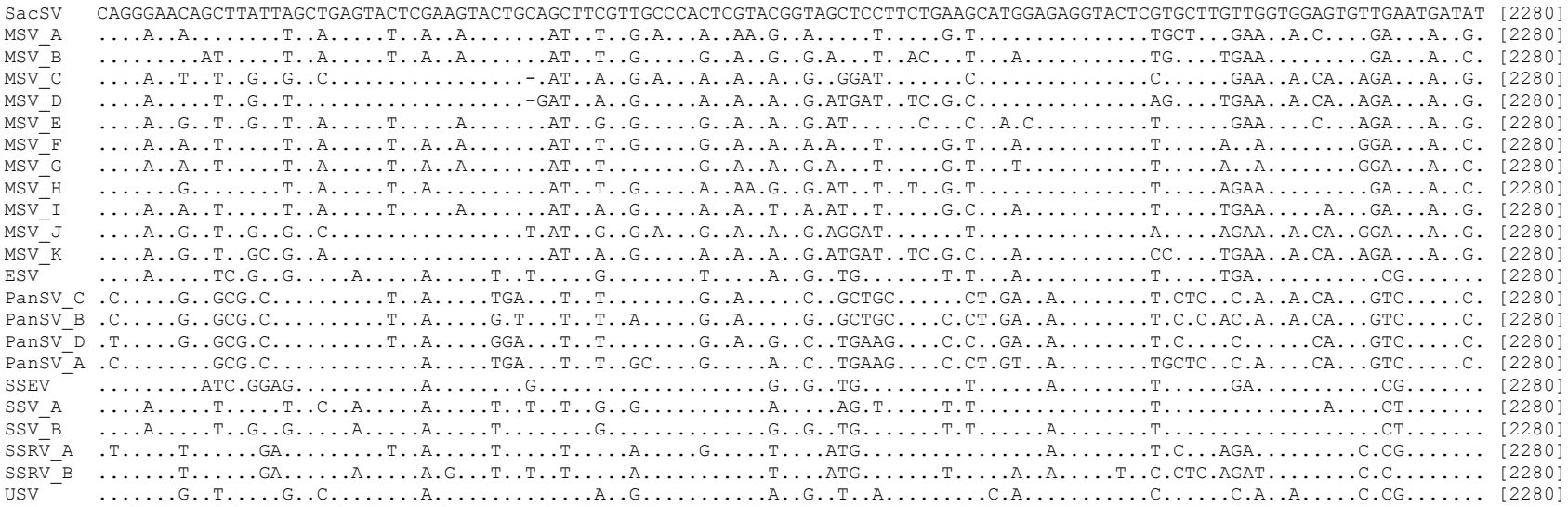
```
SacSV   CTGGCCGCGTTCCCGTACACTTGGAAGGTCGGAAGGGAGGTGTGTCATCGCTTCGTGGTGAAGCGGAGGTGGACTTTCACGATGGAGACGGACGGGCGTATCGGGAGCGATATCCCTCGG [1080]
MSV_A   ..AAAA....GG...GC...A.....A..GA.CC.....C.....................A...C.A...TTG....AC........C.....T..G..T..TTCG..........CC [1080]
MSV_B   ...AAA..T.GG...GCA..A.....A..GA.CC.....C.....................A...C.A...TTG....AC........C.....T..A..T..TTCG..C.......CT [1080]
MSV_C   ..CAAG..C.GG...ACT........A..TA.TC.....T.....................C.A...CTC.....A........T.....T..G..A..CTCG....CA..A.C. [1080]
MSV_D   ..AAAG..C.GG...ACT........A..TA.TC.....T.....................C.A...CTC.....A........C.....T..G..A..CTCG....CA..A.C. [1080]
MSV_E   ..A.TA..T.GG...ACG........A..TA.TC.....T.......C.....................C.A...CTC.....A........C.....T..A..T..TTCG..........C. [1080]
MSV_F   T....G..T.GG...ACG..C.....A..GA.CC.....T.......C..............A..CC.T...TTG..T..C........C.....T..C..T..TTCG..C..A....C. [1080]
MSV_G   ..A..T..C.GG...ACG..C........GA.CC.....A.......C..............A..CC.T...CTG.....C........CA....T..C.....TTCG..C..G....CA [1080]
MSV_H   ...AAA..T.GG...GCA..A.....A..GA.CC.....C.......C..............A..AC.....TTG....AC........C........C..T..TTCG..C.......CT [1080]
MSV_I   ....TA....GG..CACA........A..AA.CC....AC.......C..............A..C.T...CTC..T..CC.T.....C.....T..A..T...TCG.....A....C. [1080]
MSV_J   ..A.TA..T.GG...ACG........A..TA.TC.....T.....................C.A...CTC.....A........C.....T..A..T...TCG..C.......C. [1080]
MSV_K   ..AAAG..C.GG...ACT........A..TA.TC.....T.....................C.A...CTC.....A........T.....T..G..A..CTCG....CA..A.C. [1080]
ESV     T..AG...C........T..C........T.....A........C..C..G..............C.A...CTG..............CA.....A.G..T..TTCG....C....G.C [1080]
PanSV_C ..T.....A...........C........G..G........C..C...........T..T..A...C.C...TGC.....C........TA....C..G.....CTCA...G.G..A.C. [1080]
PanSV_B ..CT.A..C..T..C.....C........G..G........C..C...........T..T..A...C.C...TGC..T..C........TA....C..G.....TTCA....CT..C.C. [1080]
PanSV_D ..TT....C.....C..............T..G........C..C.....................AC.C.....G.....C........C.....T..C..T...TCG...G......C. [1080]
PanSV_A ..CT.A........C..............T..C........C..C...........T..T..A...C.C...TGC.....C........TA....C..G.....TTCA-.CG.T....C. [1080]
SSEV    ..T.TG........C.....A........TTC.....................................A....A..C..................C.....T..G..T...TCG...G.T....C. [1080]
SSV_A   T..A.A..............C........T..TC.T..A...........G.....................................AGTTA..............TTC......A....CT [1080]
SSV_B   ...A.A........C.....C........T...C.T..A...........G............A..C...AC.....T........AGT.A.......C.....T......G.....G.C [1080]
SSRV_A  T..AG...C........T..C............CC.A........C.....G..............C...................AGTCA.T..A..G..T..ATCT...G.G....CT [1080]
SSRV_B  ...AG...C........T..C............CC.A........C.....G.....................................AGTCA.T..A..C..T...TCT.....T....C. [1080]
USV     ..C..T..A........T..............T.....A.......................A...C.....................T.T......A.G.....CTC......T...A.A [1080]

SacSV   TCGACGGATTCTTGGCCGCCCTGCAAGCGGTCTATTTACTTTCACAAGTTTGCCACCGGTCTCGGTGTCAAAACGGAGTGGAAGAATCTCGCTGATGGAGGAGTTGGATCCATTAAGAAG [1200]
MSV_A   ....ATACAAG....AA...T........CAAC..C.....C........CA.G.GT..GT.G..A..G.G....C...........G.AA.G..C...........TG....CC...GA [1200]
MSV_B   ....AT.CAAG....AA...G........CAAC..C.....C........CA.G.GT..GT.G..A..G.G....C...........G.AA.G..C...........TG.G..AC...GA [1200]
MSV_C   AGC.ACC.GAG.........A..T......AACG..G....C........CA.T.GT..GT.G..A..G.GG...C...........CG..A.A..C..T........TG.G..AC...GA [1200]
MSV_D   AGC.ACC.GAG.........A..T......AACG..G....C........CA.T.G...GT.G..A..G.GG...C...........CG..A.A..C..T.........G.G..AC...G. [1200]
MSV_E   A.C.ATC.GAG.........A..T......AAC..AG....C........CA.T.G...GT.G..A..G.GG...C...........G.AA.A..C..T..G..C..TG.G..AC...GA [1200]
MSV_F   ..C.AT.CAAG....AA...G.........AAC...A....C........CA.G.GT..GT.G..A..A.G....AT..........G..A.G..C..T..T.....TAG...AC.ACGA [1200]
MSV_G   ....AT.CCAT....AAC............AGC...A....C........CA.G.GT..GT.G..A..A.G....AT..........G..A.A..C..T..C........G..AC...GA [1200]
MSV_H   ....AT.CAAG....AA...G.........AAC.........C........CA.G.GT..GT.G..A..G.G....AT..........G.AA.A..C..T........TG.G..AC...G. [1200]
MSV_I   A.C.ACC.GAG....A.T..A..T......AAC..AG....C........CA.T.G...GT.G..A..G.GG...C...........G.AA....C..T..G..C..TG.G..AC...GA [1200]
MSV_J   A.C.ATC.GAG.........A..T......AAC...G....C........CA.T.GT..GT.G..G..G.GG...C...........G.AA.A..C..T..G..C..GG.G..AC...GA [1200]
MSV_K   AGC.AC.CAAG.........A..T......AACG.GG....C........CA.T.GT..GT.G..A..G.GG...C...........CG..A.A..C..C.........TG.G..AC...GA [1200]
ESV     C.T..CAG..GC.....C..G...CGCAA.AAC.........C..........AAG....CT.A..G..G..................CACTA.A.GC....AC..C..TGA.......... [1200]
PanSV_C ..C.ACACAG.G.....C..A..T...AA.GAC..C.....C........CAG...G..CT.G..C..G..G..............CG.TA.A..C..GAAG.AC..CG....C...... [1200]
PanSV_B ..C.AC.TGG.......A.........AA.GAC..C.....C........CTG...G..CT.G..C..G..G..............CG.TA.A..C..GAAG.AC..CG....C...... [1200]
PanSV_D G.T.AT.CC...........G........C..C..A.....C........CTG...T..A.....A..G..G..............CG.TA.A..C..GAAG..C..CG........... [1200]
PanSV_A G.C.ATACCG..........T......AA.GAC..C.....C........CTG...G..A.....C..G..G..............CG.TA.A..C..GAAG..C..CG.T......... [1200]
SSEV    G...ATACC..A........T..T......G.G.........C..T.....CTT...T..C..G.....G..G.....................A.A..C..T..T..G...G.G......... [1200]
SSV_A   AGT..CAGC.G......C..G...CGTAA.AAC.........C..........T......CT.A.....T..................CACTA.C.GC..T.A...C..TGA.......... [1200]
SSV_B   AGC..CAGC.G......C..G...CGTA.AAA.........C..........T......CT.A..G..A..................CACTA.C.GC..T.AT..G..TGA.......... [1200]
SSRV_A  G....T.CCGTG.....C..G...CG.AA.AAC..A.....C..........T......A..A.....G..................ACAA.C.GC..T.A...C...GA.......... [1200]
SSRV_B  G...AT.CAGTG.....C..G...CG.AA.AAC..A.....C..........T......A..A..G..A..................ACAA.C.GC..T.AT..C..TGA.......... [1200]
USV     A........G.....................AG...C.....C..........T......A........A.....C...........CG.A.....C.....C..A..GG........... [1200]
```


Coat protein stop codon
Virion-sense gene polyadenylation signal
Complementary-sense gene polyadenylation signal[5]
Complimentary-strand origin of replication


Sequence conserved in PanSV
Replication-associated protein stop codon
Inverted repeat
Loop sequence

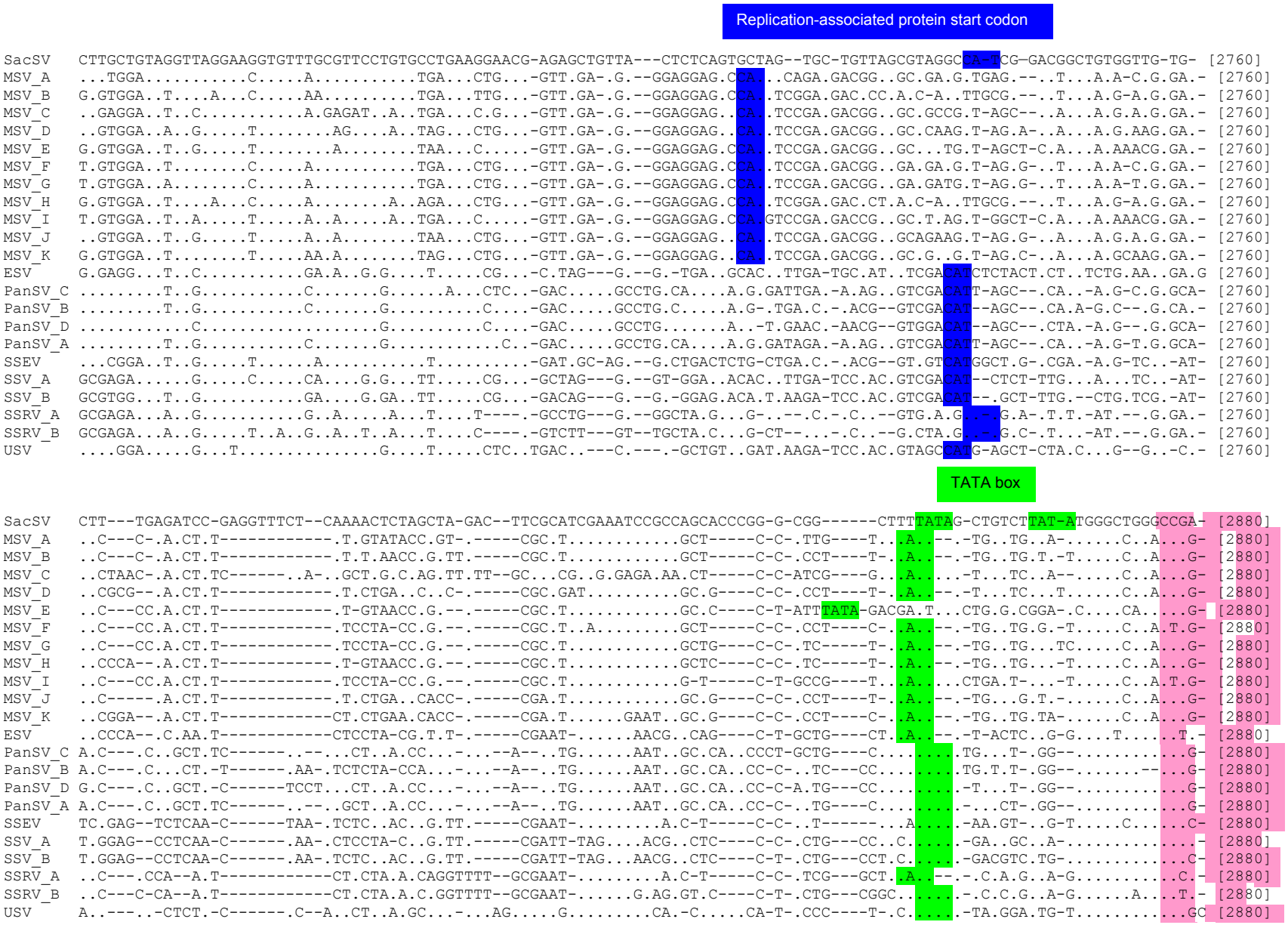

RepA stop codon


Intron acceptor AG
Alternative branch site
Branch site
Intron
T-Tracts
Intron donor GT


SacSV CAGGGAACAGCTTATTAGCTGAGTACTCGAAGTACTGCAGCTTCGTTGCCCACTCGTACGGTAGCTCCTTCTGAAGCATGGAGAGGTACTCGTGCTTGTTGGTGGAGTGTTGAATGATAT [2280]
MSV_A ....A..A........T..A.....T..A..A.......AT..T..G.A...A..AA.G..A.....T.....G.T..............TGCT...GAA..A.C....GA...A..G. [2280]
MSV_B .........AT.....T..A.....T..A..A.......AT..T..G.....G..A..G..G.A...T..AC...T...A...........TG....TGAA.........GA...A..C. [2280]
MSV_C ....A..T..T..G..G..C..................-.AT..A..G.A...A..A..A..G..GGAT.......C..............C......GAA..A.CA..AGA...A..G. [2280]
MSV_D ....A.....T..G..T.....................-GAT..A..G.....A..A..A..G.ATGAT..TC.G.C..............AG....TGAA..A.CA..AGA...A..G. [2280]
MSV_E ....A..G..T..G..T..A.....T.....A.......AT..G..G.....G..A..A..G.AT......C...C..A.C..........T......GAA....C...AGA...A..G. [2280]
MSV_F ....A..A..T.....T..A.....T..A..A.......AT..T..G.....G..A..A..A.A...T.....G.T...A...........T.....A..A........GGA...A..C. [2280]
MSV_G ....A..A..T.....T..A.....T..A..A.......AT..T........G..A..A..G.A...T.....G.T...T...........T.....A..A........GGA...A..C. [2280]
MSV_H .......G........T..A.....T..A..........AT..T..G.....A..AA.G..G.AT..T..T..G.T..............T.....AGAA.........GA...A..C. [2280]
MSV_I ....A..A..T.....T..A.....T.....A.......AT..A..G.....A..A..T..A.AT..T.....G.C...A...........T.....TGAA.....A...GA...A..G. [2280]
MSV_J ....A..G..T..G..G..C..................T.AT..G..G.A...G..A..A..G.AGGAT.......T..............A.....AGAA..A.CA..GGA...A..G. [2280]
MSV_K ....A..G..T..GC.G..A...................AT..A..G.....A..A..A..G.ATGAT..TC.G.C...A...........CC....TGAA..A.CA..AGA...A..G. [2280]
ESV ....A.....TC.G..G.....A.....A.....T..T.....G........T.....A..G..TG.......T.T...A...........T.....TGA............CG....... [2280]
PanSV_C .C.....G..GCG.C..........T..A.....TGA...T..T........G..A.....C..GCTGC......CT.GA..A........T.CTC..C.A..A.CA...GTC.....C. [2280]
PanSV_B .C.....G..GCG.C..........T..A.....G.T...T..T..A.....G..A.....G..GCTGC....C.CT.GA..A........T.C.C.AC.A..A.CA...GTC.....C. [2280]
PanSV_D .T.....G..GCG.C..........T..A.....GGA...T..T........G..A..G..C..TGAAG....C.C..GA..A........T.C....C.......CA...GTC.....C. [2280]
PanSV_A .C........GCG.C..............A.....TGA...T..T..GC....G.....A..C..TGAAG....C.CT.GT..A........TGCTC..C.A....CA...GTC.....C. [2280]
SSEV .........ATC.GGAG...........A........G.....................G..G..TG.........T......A........T......GA............CG....... [2280]
SSV_A ....A.....T.....T..C..A.....A.....T..T..T..G..G...........A.....AG.T.....T.T..............T..............A....CT....... [2280]
SSV_B ....A.....T..G..G.....A.....A.....T........G..............G..G..TG.......T.T......A........T...................CT....... [2280]
SSRV_A .T.....T......GA.........T..A.....T.....T.....A.....G.....T....ATG..............A........T.C...AGA.........C.CG....... [2280]
SSRV_B .......T......GA......A.....A.G...T..T..T.....A...........T....ATG.......T.....A..A.....T..C.CTC.AGAT........C.C........ [2280]
USV .......G..T.....G..C........A..............A..G...........A..G..T..A.........C.A...........C......C.A..A.....C.CG....... [2280]

SacSV CGCGAACGATATCATCCTTTGAGGCGCGAGTGTTA---GA---GCTCTCACCGATCTGAGGAACGAAGGGCTTCTTCCGTGGAACGAAAGTACCCTTCTCCCACTGACACAATGGGTCCT [2400]
MSV_A .T..C.TT..T.....T..A..A.GCTTTT.T.CC---TTTACCTCTGA.T.AGAT.TTCCT.G......GGA.....TA....T.........TC....AA..AC.GC..GA..T.... [2400]
MSV_B .T.TC.TT..T.....T..G..A.GTTTCT.T.C.GAATTTCCCT.GGA.GGA---.TTCCTTG....CTTGA.....TA....T.........TC....AA..AC.GC..GA..T.... [2400]
MSV_C .T..C.TT..T.....T..G..T.GTTTCT.A.CCGAATTTCCCT.TGAGGA.---.TTCCT.G...T.AT.......TA....T...T.....TC....AA..AAGGC..GG..T.... [2400]
MSV_D .TT.C.TT..T.....T..A..T.GTTTTT.T.CTGAATTTCCCT.TGAGGA.---.TTCCT.G...T.TT.......TA.............TC....GA..AA.GC..GA..T.... [2400]
MSV_E GTGCC.TT..T.....T..GTTT.GTTTTT.T.CG---..---AGGGGA.GTA---.TTCCTTG....CATGA......A....T.........TC....AA.TACTGC..G...T.... [2400]
MSV_F .T.TC.TT..T.....T..GTT..GTTTCC.AGA.---TTTCCCTCTGA.T.AGAA.TTCCT.G...T..AGA.....TA..G.T.........TC....AA.TAA.GCT.GA..T.... [2400]
MSV_G .T.TC.TT..T.....T....TA.GTTTCT.TGA.---CTTCCCTCTGA.T.AGAT.TTCCT.G......GGA.....TA....T.........TC....AA.TAA.GC..GA..T.... [2400]
MSV_H .T..C.TT..T.....T..G.TA.GTTTCT.TGA.---TTCCCCTCTGA.T.AGAT.TTCCT.G......TGA.....TA....T.........TC....AA..AAGGC..GA..T.... [2400]
MSV_I .T..C.TT..T.....T..ACTT.GTTTCT.T.C.GAATTTCCCT.GGA.GA.---.TTCCT.G....CATGA.....TA....T.........TC....AT.TACCGC...G..T.... [2400]
MSV_J .T..C.TT..T.....T.....A.GTTTT..T.C.GAATTTCCCT.TGA.GA.---.TTCCT.G.....TT.......TA....T.........TC....GA..AAGGC..G...T.... [2400]
MSV_K .TT.C.TT..T.....T..A....GTTTCT.T.C.GAATTACCCT.TGAGGA.---.TTCCT.G..CA.AT.......TA....T.........TC....GA..AAGGC..GA..T.... [2400]
ESV .T.TC..............G..TTGCTT....G.G---TTCCCCTCTGAGGTAGA-GCC--..G.....TT.......T...T.T...T.....TC........ACCTT..G.....T.. [2400]
PanSV_C .T.TC.T..CT........A.TA.GCTTCT.C.C.---..---A.....CTTACCTG.C...CT.....A........T.....T.T.T.....T..........TGTC.TTC..A.... [2400]
PanSV_B .T.TC.T..CT........G.TA.GTTTCT.C.C.---.C---TGG...CTTACCTG....G......CT........T...G.T.T.T.....T..........TGTC.TTG....... [2400]
PanSV_D .T.TC.T..CT........GCTA.GCTTCT.CCC.---..---.T.T..CTTACCTG....C........A........T...T.T.T.T.....T..........TGTC.TTG....... [2400]
PanSV_A .CTTC.T..CC........A..A.GCTTCT.C.C.---..---AT.T..CTTACCTG...........CT........T...T.T.T.......T..........TGTC.TTG..A.... [2400]
SSEV .T.TC...........T....T..GTTT..GT.GC---CTATCCTCAGATGAAG.-G.T--.G.A....AT.....G.....G.T.........T.........T..TTTG.TC....TT. [2400]
SSV_A .T.TT.............G..TTGCTTCT..GA.---TTTGGCTCAGTCGAA---.CTCCT.G.....AGG......TA..T.T...C.....TC.......G.ATGTA.CC....T.. [2400]
SSV_B .T.TG...........T.....TTGTTTTT..GA.---TTTGGCTCAGTCGAA---.CTCCCTG....CTAG......TG..T.T...T.....T........AAT..TT.CC..C.T.. [2400]
SSRV_A .T.T.............G..T.GTTTG.AA..T---TTGCTCTCAGAGGAAGA-G..--T..A...CAT.GT..G.......TA..T.....T........T.T.GCG.T...A.TT. [2400]
SSRV_B .T.T...........T..G....GTTTG.AA..C---TTGCTCTCAGAGGAAGA-G.T--C......CTT.G...G.......T...C.....T.............T.C.....T.. [2400]
USV .T.T.................TT.GTTAG....G-----TGCTCA..G.CTT.G.-G-.........A.AT.....T.......T...T.....TC.......TA.TGCG.T...A..T. [2400]

```
SacSV   TGAGTATGTACTCTCTTACCCTGTCTACCGATTTAGCAGACTGGATGTTGGGGTGGTACTCATTTACGTCGAAGAACCGCGGGTCTGAAGTCGTCACGGGGCGGACGCTCTGTGCCAGGG [2520]
MSV_A   ....A.....A..C..C..T....TA..T..C..G...CT...A..A..T.....AA..C......TA..A.......TT.A...A..TA..C.T..C..CTTCT.TG....AAG..AT. [2520]
MSV_B   ....A.....A..C..G...T...T...T.....T...CT...A.....T......A..C.T.CG.TA..A.......TT.A...A.TGA.TC.T..C..TTTCT.TG.T..GAGT.AT. [2520]
MSV_C   ..GTG.....A....G....T...TGGGGCTGA.T...CT...A...........A.AC...CA.T...........TT.A...G.T....C.A..T..TTTCT.CG.....ATT.AT. [2520]
MSV_D   ....G.....A.........T...TAGGGCTCA.G..GCT...A.....T......A.AC...C..T...........TG.A...G.TG..TC....TT.TTTAT.CG.....ATT.AT. [2520]
MSV_E   ....A.....G..G.G....T...T...G..C......CT...A.....T.....AA.TC.T.C..T......A....TG.CA..G.TC...C....T..TTTAT.CG.....AGT.AT. [2520]
MSV_F   ....A.....G....GA...T...T...T..C..T...CT...A.....T......A.GC.C..G.TA..A..A....TT.A...A.TGA..C.T..C..TTT.T.CG....GA.T.AT. [2520]
MSV_G   ....A.....G....GA...T...TA..T..C..T...CT...A.....T......A..C.T..G.TA..A.......TT.....A.TT..TC.T..C..TTTCT.CG....GA.T.AT. [2520]
MSV_H   ..........A.....G..TT...T...A.....G...CT...A.....A......A..C......T...A.......TT.A...A.TGA.TC.T..C..TTTCT.TG....GATT.AT. [2520]
MSV_I   ....A.....A..A.G....T...T...G.........CT...A.....T......A..C.C.CA.TA..A..A....TG.AA..G.TC..TC.T..T..TTTAT.CG.....AGT.AT. [2520]
MSV_J   ..........T.....G...T...T...GCT..GC...CT...A...........A.GC...CA.TA..........TG..A..G.T...TC.T..T..TTTTT.CG.....A.T.AT. [2520]
MSV_K   ....G.....A..C......T...TG..GCTCA.G...CT...A...........A.AC...CG.T...........TG.A...G..G..TC.A..T..TTT.T.CG....AATT.ATA [2520]
ESV     .C.A..C...A..C......T...TAG.ACTC..T....CT..T.............G..T.CG.T...........TTTCA..ATGG..GTAG..A..CTT.........A..T.A.. [2520]
PanSV_C .C........G..A.GC...T....AGTGCTC..G..GCT...T.....C..A......CG..C..T...A...T...T.TC...ACG...A..G..A..CTT..T..A...GAGA..A. [2520]
PanSV_B .C...........A.GC...T....AGTGCTC..G..GCT...T.....CCCA......CG..C..T...A...T...T.TC...ACG...A..GCAA..CTT..T..AT..CAGG..T. [2520]
PanSV_D .C........G..A.GC...T....AGTGCTC..G..GCT...T........A......CG..C..T...A...T...T.TCA...CTT..T..G..T..CTT..T..AT..CAGG..A. [2520]
PanSV_A .C...........A.G....T....AGTGCTC..G..GCT...T..A..C..A..A...CG..C..T...A....-..T.T-...ACG...A..G..A..CTT..T..A...GAGA..A. [2520]
SSEV    ....C.C...T..CT.G.........G.G..C..G...CT...A.....A..A...A.A..C.CA.T...A.......TA....TC.T...TTG...A..CTT.........A..T.A.. [2520]
SSV_A   .C.TGGC...TG....G...T...TGG.ACTC..T...CT...A........A...A.GC...CA.T.........G.CG.A...A.TT..GTAG..A..TTTTG.A.....A..G..TA [2520]
SSV_B   .CGTG.....TG..T.G.T.T...TAG.ACTC..TC..CT..............A.GG.T..CA.T.........GGC...A..C.TG..GTAG..A..CTT...A.....G..G..TA [2520]
SSRV_A  .C..G.....T..C..G..TT...TGG.ACTC..T...CT...A.....A......A.AC.C.CA.T...........CTT...TC.TC..TTGT..A..C.T..........A.T.A.. [2520]
SSRV_B  .C........T.....G...T...TAG.GCT.......CT...A.....A......A.AC.C.C..T...A.......C.T...TC.TT..GTGT..A..CTT...........T.A.. [2520]
USV     .C........GG.........T.......A..C.....T..T........C..A..A.GA..T.C..T.........T.TT.AA.T.CTG.....G..A..CTT...........A..A. [2520]

SacSV   CATGGCAGTGCCATGTCCCATCTTGATGAGCTTCTCTAGCCACCAGGATGTATGCCGGAGTCC-ATGGAG-CTAGCTTGCTCCATAGGCTCAGACCCAAGATCTCGGGATCAAGGCTACA [2640]
MSV_A   ....TA.A...A.AC.T.......T...T..C....GG..ACAT..A..A..CTTG...A...-.AC..A-.G.CGAGCTC...G.TCA..T...AGGC...T..A....TTTCTGG... [2640]
MSV_B   ....TA.A...ATATCT..G....TG..T..C...TGG..ACAT.T...A..CTT...G....-.AC..C-.A.CGAGCTC...A.TCA..T...AG.C...T..T....TTTCTGG... [2640]
MSV_C   ....TA.ACATTGAT.T......CT...C..C..C.GG..GCAT.T......CAGA.......-.AT..C-TACAGAGATC...G.TC...T...TG.T...T..T..C.GTTC.GG... [2640]
MSV_D   ....TA.A...ATAT.T.....C.C...T..C...TGG..ACAT.T......CAGA...T...-.AC..C-.A.CGAGATC...G.TC...T...TG.T...T..A....GTTCTGG... [2640]
MSV_E   ....TA.....ATAT.T.....AC....T..C....GG..ACAT.C......CAGA...T...-.AC...-.G.CGAGATC......TT..T...TG.T...T..T....TTTCTGG... [2640]
MSV_F   ....TA.A...A.AT.T......CT...T..C....GG..ACAT.T......CTTA..GT...-.AC..C-.A.CGAGCTC...G.TTA..T...AG.C...T..T....TTTCTGG... [2640]
MSV_G   ....TA.A...AGAT.T......CT...T..C....GG..ACAT.T......CTTA..GT...-.GT..C-.A.CGAGCTC...G.TCA..T...AG.C...T..T...GTTTCTGG... [2640]
MSV_H   ....TAGA...ATATCT.......TG..T..C....GG..ACAT.T......CTTA..GT...-.AC..C-.G.CGAGCTC...G.T.A..T...AG.C...T..T....TTTCTGG... [2640]
MSV_I   ....TA.A...ATAT.T.....AC....T..C....GG..ACAT........CAGA...T...-.AT...-.G.CGAGATC......TT..T...TG.C...T..T....TTTCAGG... [2640]
MSV_J   ....TA.A...ATAT.T......CT......C....GG..ACAT.T......CAGA...T...-.AC..C-...CGAGATC...A.TC...T...TG.T...T..T....TTTCTGG... [2640]
MSV_K   ....AAGA..TA.AT.T.....C.C...T..C....GG..ACAT.T......CAGA..CT...-.AC..C-...CGAGCTC...G..TT..T...TG.C...T..T....TTTCTGG... [2640]
ESV     ....CAGA.......A...................C...CT.....C.........A.......-.-T..CT...AGAGA.........CAC......T.C..CT..A..C..T.A.G.G.. [2640]
PanSV_C ...................G...GCG..T.TC.....GA....T..T.....-AG.T.GC..AA.GTCCT-TAGTGAGC..GA.G..ATG.TC...G.T..A...T..T..T..TGG... [2640]
PanSV_B ...................G...GTG..T.TC..C..GA.T..T.A......-TG.T.GC..AT..TCCC-TAGTGAGT..GA.G..ATG.TC...G.T..A...T..T..T..TGG... [2640]
PanSV_D .................T......GTG..T.TC......A.T.....T.....-.G.T.GT..AA.GTCCC-TAGTGAGC..GA.C..ATG.TC...G.T..A...T..T..T..TGG... [2640]
PanSV_A .................T..G.....G....TC.....CA.T...........-TG.T.GC..AA.GTCCT-TAGTGAGC..GA.G..ATG.TC...G.T..A...A..T..T..TGG... [2640]
SSEV    .....ATA.....G.AT.....AGC.........C.GGCT...G.C......A..T...T...-.GT..C-...TGAGA.....G..ATGG..C..T.CAGCT..A..T..T..ATGG.. [2640]
SSV_A   .G...AT...AT.CCCT..G....C...T.TC..C..GA.AGAT....CA..CA.T..T....-.GT.T.-.A.TGAG........T.TGT.....T.CAGCT..A..T..T.ATGG... [2640]
SSV_B   .G...ATA..AT..CCG..G..C.CG..T.TC.....GA.AGAG..T........G...T...-.GT...-AG.TGAG.T....G.T.A.......T.CAGCT..A..G..T...T.G.. [2640]
SSRV_A  .....AG.......TCT......GCG..T.....C..GCT...T..A.....C......C...-.GT..C-TA.TGAGCTG......TA.........CTGC...T..G..C..T..G.. [2640]
SSRV_B  .....AG.......TCT.....AGC......C....GGCTT..T..A.....A..A...T...-.GT..C-.A.TGAGCTC...G...TGG.......CTGC...T..C..C..A..G.. [2640]
USV     ...................G.....G..T..C....G.CA...T.T............TTC..-......-...T.AA..........G.G.....A.T.GCT..A..C..T..AGGG.. [2640]
```


Replication-associated protein start codon
TATA box

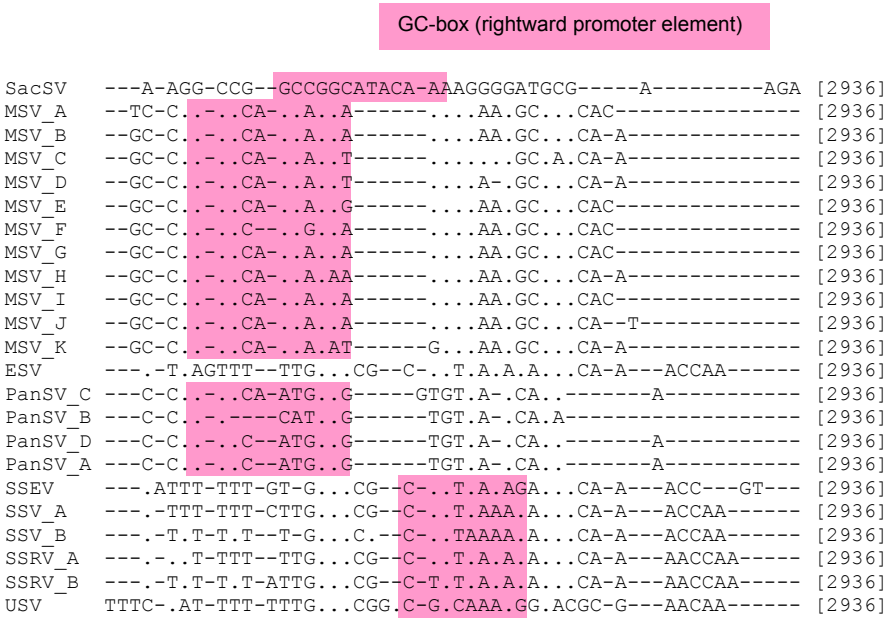


Figure 2.4: Genome sequence annotations of SacSV [ZA-Emp-T1-2008] together with a selection of major strain variants from other publicly available African streak virus species (GenBank accession numbers are provided in Table 2.1). Sequences either known or believed to have some role in mastrevirus replication and transcription are marked together with a corresponding label on the nucleotide sequence alignments. To highlight differences between the sequences, wherever nucleotides in a particular alignment column are identical to that of SacSV [ZA-Emp-T1-2008], they are replaced with a "." character. Gaps introduced to optimise the alignment are indicated with a "-" character.Full genome annotation of SacSV, showing key gene regions.

2.4.2 Evolution

Identification of the various reading frames allowed us to compare sequence identity at various sites on the genome (ORFs Table 2.1). The full genome of SacSV was most similar (sharing ~65.4% identity) with that of USV - a streak virus species that has only currently been found infecting *Urochloa deflexa* in Nigeria. Therefore, based on the ICTV 75% identity species demarcation threshold (Stanley *et al., 2005*), we propose the new isolate be called **Sac**charum **S**treak **V**irus (SacSV; full name SacSV-[ZA-Emp-T1-2008]). Although the overall genome of SacSV is most similar to Urochloa other reading frames show different relationships. The full genome of SSEV (SSEV-[EG-Egypt]) and PanSV-D (PanSV-D-[NG-Ifo-g91-2006]) show almost equally relation (~55%) to SacSV, however the *MP* and *Rep* of SSEV are much more closely related to SacSV (Table 2.1).

Whereas we determined that SacSV-like viruses have most probably contributed genetic material to the evolution of other streak virus species, it is itself not obviously an inter-species recombinant. Most notably, our analysis revealed that a SacSV-like virus is the likely parental donor of a previously detected ~50 nucleotide recombination tract within the MSV strain F SIR (Varsani *et al.,* 2008a); see supplementary .rdp project file for details) . This tree (Figure 2.5) clearly confirms that USV is the closest currently sampled relative of SacSV.

						Percentage pairwise sequence identity of the full genone[A] *mp*, *cp* and *rep* (full length) genes[B] and proteins products[C] of various African streak viruses to SacSV						
GenBank accession #	**Name**	**Host**	**Longitude**	**Latitude**	**Short name**	**Full genome[A]**	***mp*[B]**	***mp*[C]**	***cp*[B]**	***cp*[C]**	***rep*[B]**	***rep*[C]**
GQ273988	SacSV-[ZA-Emp-T1-2007]	Sugarcane	31.88572	-28.7373	SacSV	-	-	-	-	-	-	-
EU445697	USV-[NG-Ipe-g226-2007]	*Urochloa deflexa*	4.45	7.51667	USV	65.4	69.4	63.1	75.7	85.7	68.7	68.9
EU224265	PanSV-D-[NG-Ifo-g91-2006]	*Urochloa maxima*	5.776853	6.901831	PanSV_D	57.4	57.3	31.6	64.5	69.3	63.3	60.4
AF239159	SSEV-[EG-Egypt]	Sugarcane			SSEV	57.1	64.4	47.3	68.3	69.9	64.8	64.6
L396381	PanSV-A-[ZA-Kar-1994]	*Panicum maximum*	31.18425	-25.4951	PanSV_A	55.9	55.0	33.5	59.9	68.2	63.9	61.3
EU244915	ESV-[ZM-Gur-g186-2007]	*Eragrostis curvula*	31.1021	-17.8101	ESV_ZM	55.0	56.1	30.7	63.5	60.1	59.9	61.3
AF072672	SSRV-[RE-Reu]	Sugarcane			SSRV_A	54.8	55.3	40.3	65.4	69.5	55.9	58.4
EU244916	SSRV-B-[ZM-Nya-g177-2006]	*Paspalum conjugatum*	32.9715	-18.3213	SSRV_B	53.8	52.0	37.8	66.7	69.5	56.8	60.6
EU224264	PanSV-C-[ZM-NGur-g169-2006]	*Urochloa plantaginea*	30.8402	-17.5216	PanSV_C	53.6	55.5	39.5	56.5	74.2	60.9	57.0
X60168	PanSV-B-[KE-Ken-1991]	*Panicum maximum*			PanSV_B	52.2	55.8	16.8	56.0	65.8	58.1	56.1
M82918	SSV-A-[ZA-SN]	Sugarcane			SSV_A	51.3	53.1	43.6	65.8	70.4	52.6	55.2
EU628641	MSV-J-[ZW-Mic24-1987]	*Pennisetum sp.*	30.96333	-17.8761	MSV_J	49.4	44.2	30.7	54.8	57.0	56.2	55.8
Y00514	MSV-A-[ZA-SA-1986]	Maize	26.82875	-26.7688	MSV_A	48.2	51.3	24.5	55.7	57.6	54.5	57.6
EU244914	SSV-B-[RE-Pie-R5-2006]	*Cenchrus myosuroides*	55.4817	-21.3143	SSV_B	48.2	54.2	46.8	62.1	72.0	51.3	54.8
EU628638	MSV-H-[NG-Lag-g74-2007]	*Setaria barbata*	4.666667	8.916667	MSV_H	47.3	39.7	32.7	53.0	59.5	55.4	55.8
AF329889	MSV-D-ZA-[Raw-1998]		19.74933	-33.7435	MSV_D	46.3	42.6	30.7	52.3	59.5	53.4	53.1
EU628639	MSV-I-[ZA-NewA-g217-2007]	*Digitaria ciliaris*	30.89359	-29.8126	MSV_I	46.3	39.0	30.7	52.6	60.1	53.2	57.1
EU628626	MSV-E-[ZA-MitA-g125-2006]	*Digitaria ciliaris*	31.00939	-29.8259	MSV_E	45.8	44.2	26.6	52.8	58.2	53.5	54.1
EU628631	MSV-G-[TD-Mic24-1987]	*Digitaria sp.*	-7.8882	12.18787	MSV_G	45.8	43.0	20.1	52.0	57.6	54.8	56.2
EU628597	MSV-B-[ZA-PlaB-g27-2006]		19.92809	-33.6634	MSV_B	45.7	49.9	30.7	54.4	57.0	52.1	54.0
AF007881	MSV-C-[ZA-Set-1998]	*Setaria sp*	31.03495	-29.697	MSV_C	45.3	39.1	30.7	51.8	58.2	51.5	53.1
EU628643	MSV-K-[UG-BusD-2005]	*Eustachys petraea*	30.40586	0.458333	MSV_K	45.3	36.7	49.8	51.5	69.9	52.4	53.1
EU628629	MSV-F-[NG-IntB-g88-2007]	*Urochloa maxima*	3.898865	7.406774	MSV_F	44.0	43.6	20.1	49.2	58.8	53.5	56.2

Table 2.1: Percentage pairwise distances (calculated in MEGA 4 (Tamura *et al., 2007*) with pairwise deletion of gaps) of the full genome[A] and the *cp*, *mp* and full length *rep* nucleotide (***mp*[B], *cp*[B], *rep*[B]**) and amino sequences (***mp*[C], *cp*[C], *rep*[C])**

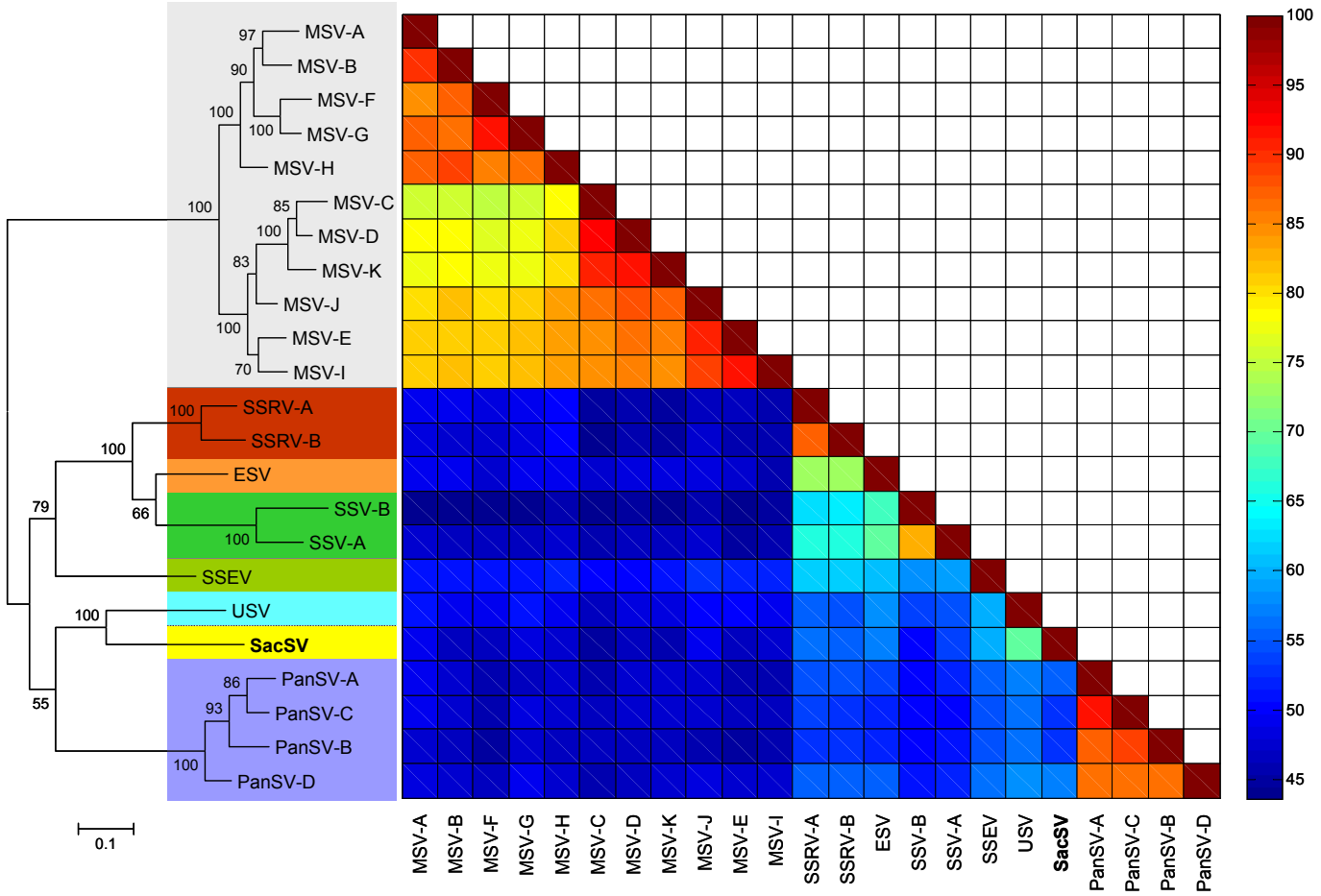


Figure 2.5: Full Genome: A heat plot and phylogenetic tree showing the level of relatedness across the full genome of a selection of mastrevirus species. Trees were drawn with PHYML using the GTR+I+G4 model. Numbers on each branch indicate the percentage of 100 full maximum likelihood bootstrap replicates that support the existence of each branch. Heat plots display precent identity based on pairwise deletion matrices.

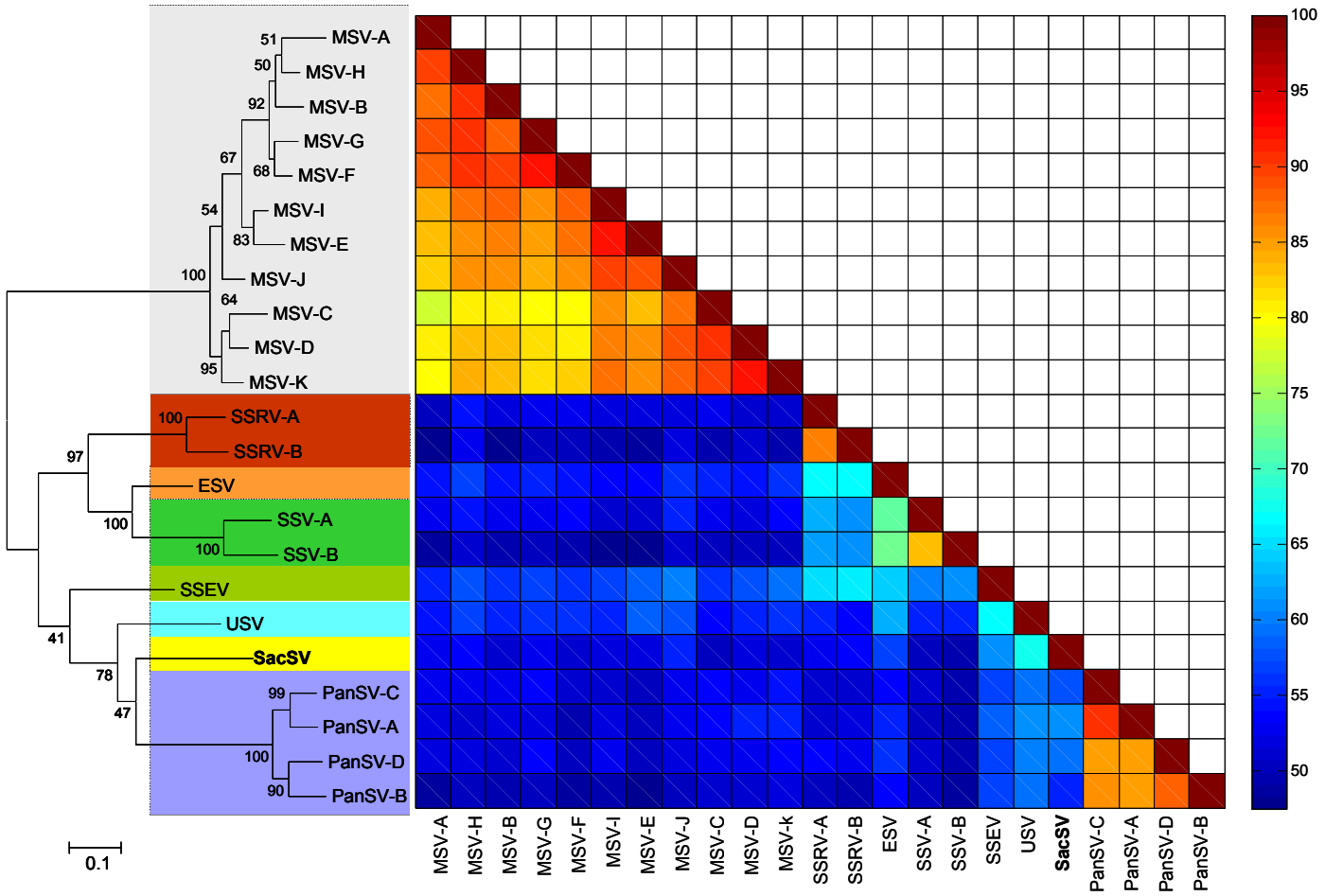


Figure 2.6: Replication Protein: A heat plot and phylogenetic tree showing the level of relatedness within the replication protein of a selection of mastrevirus species. Trees were drawn with PHYML using the GTR+I+G4 model. Numbers on each branch indicate the percentage of 100 full maximum likelihood bootstrap replicates that support the existence of each branch. Heat plots display precent identity based on pairwise deletion matrices.

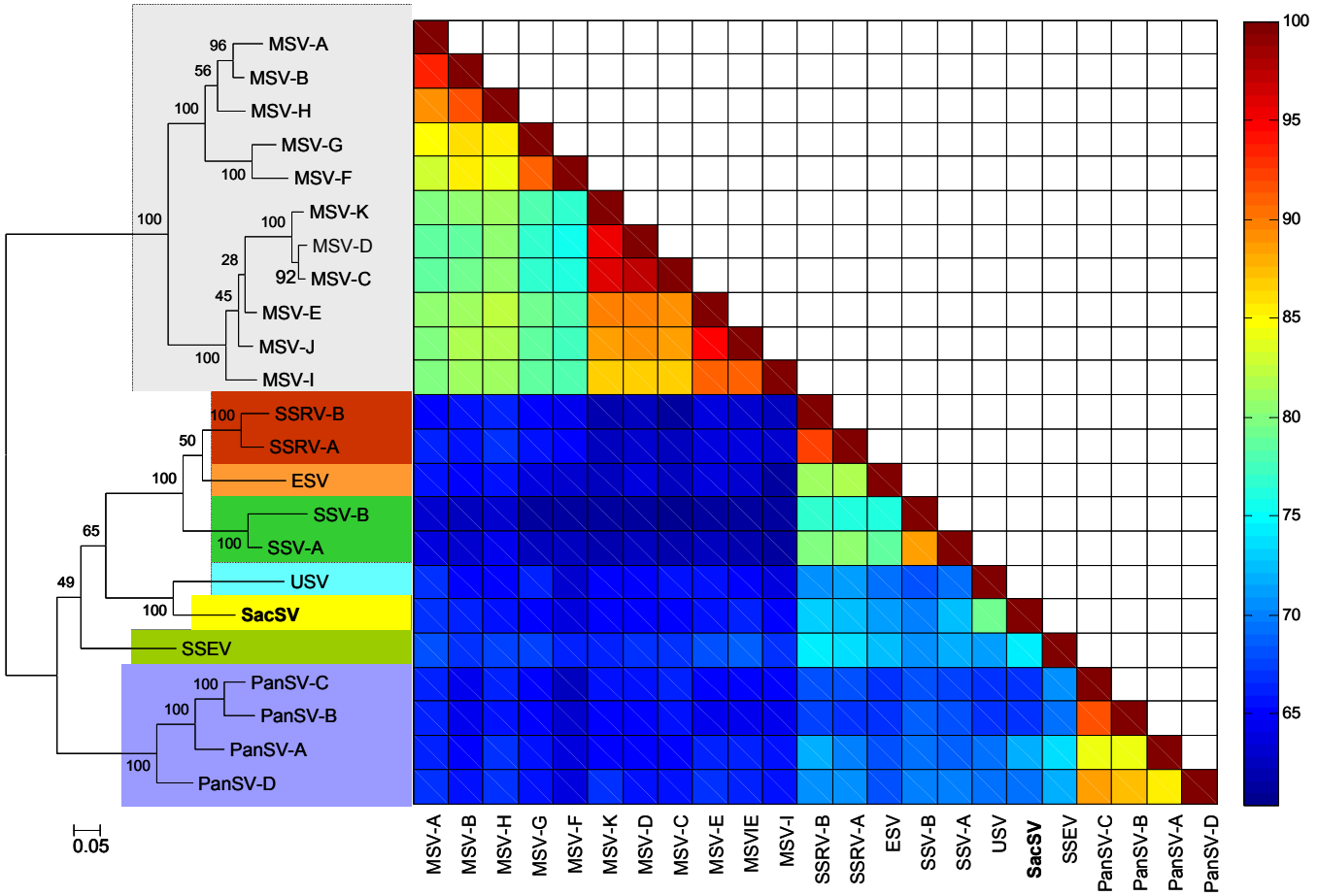


Figure 2.7: Coat Protein: A heat plot and phylogenetic tree showing the level of relatedness within the coat protein of a selection of mastrevirus species. Trees were drawn with PHYML using the GTR+I+G4 model. Numbers on each branch indicate the percentage of 100 full maximum likelihood bootstrap replicates that support the existence of each branch. Heat plots display precent identity based on pairwise deletion matrices.

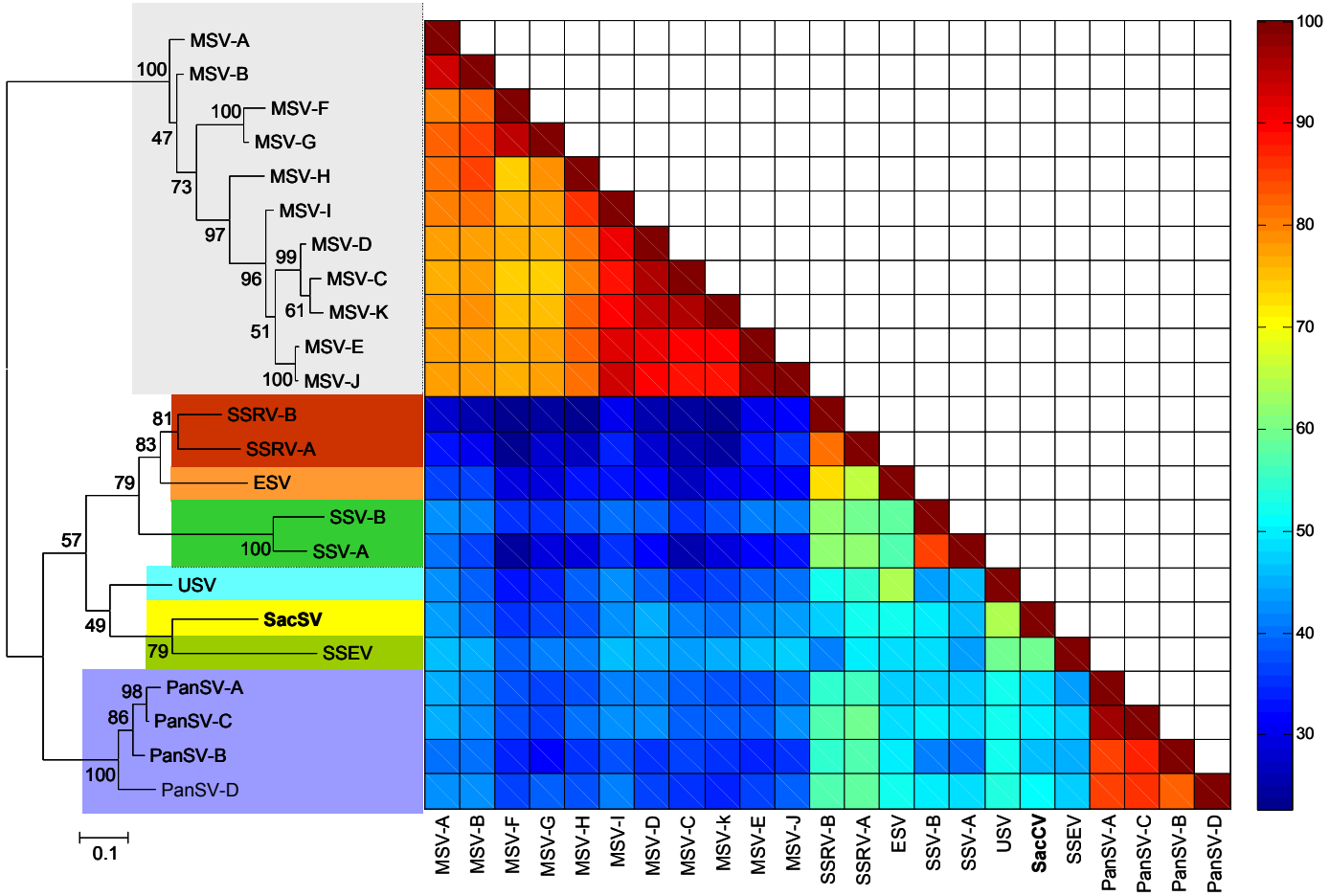


Figure 2.8: Movement Protein: A heat plot and phylogenetic tree showing the level of relatedness within the movement protein of a selection of mastrevirus species. Trees were drawn with PHYML using the GTR+I+G4 model. Numbers on each branch indicate the percentage of 100 full maximum likelihood bootstrap replicates that support the existence of each branch. Heat plots display precent identity based on pairwise deletion matrices.

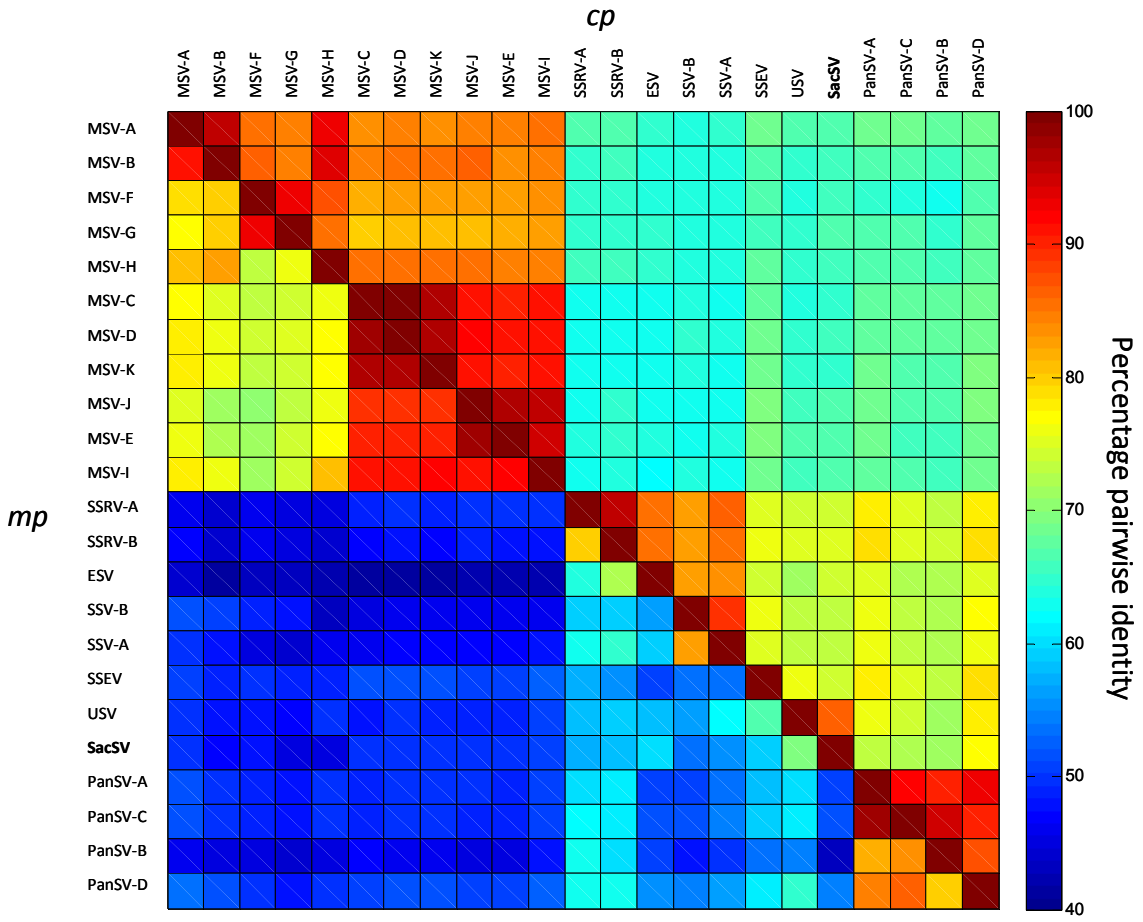


Figure 2.9: A comparison of relatedness between the MP and CP regions of a selection of mastreviruses. The order of species is based upon the trees above. Plots were constructed based on pairwise distance matrices, showing percent pairwise identity.

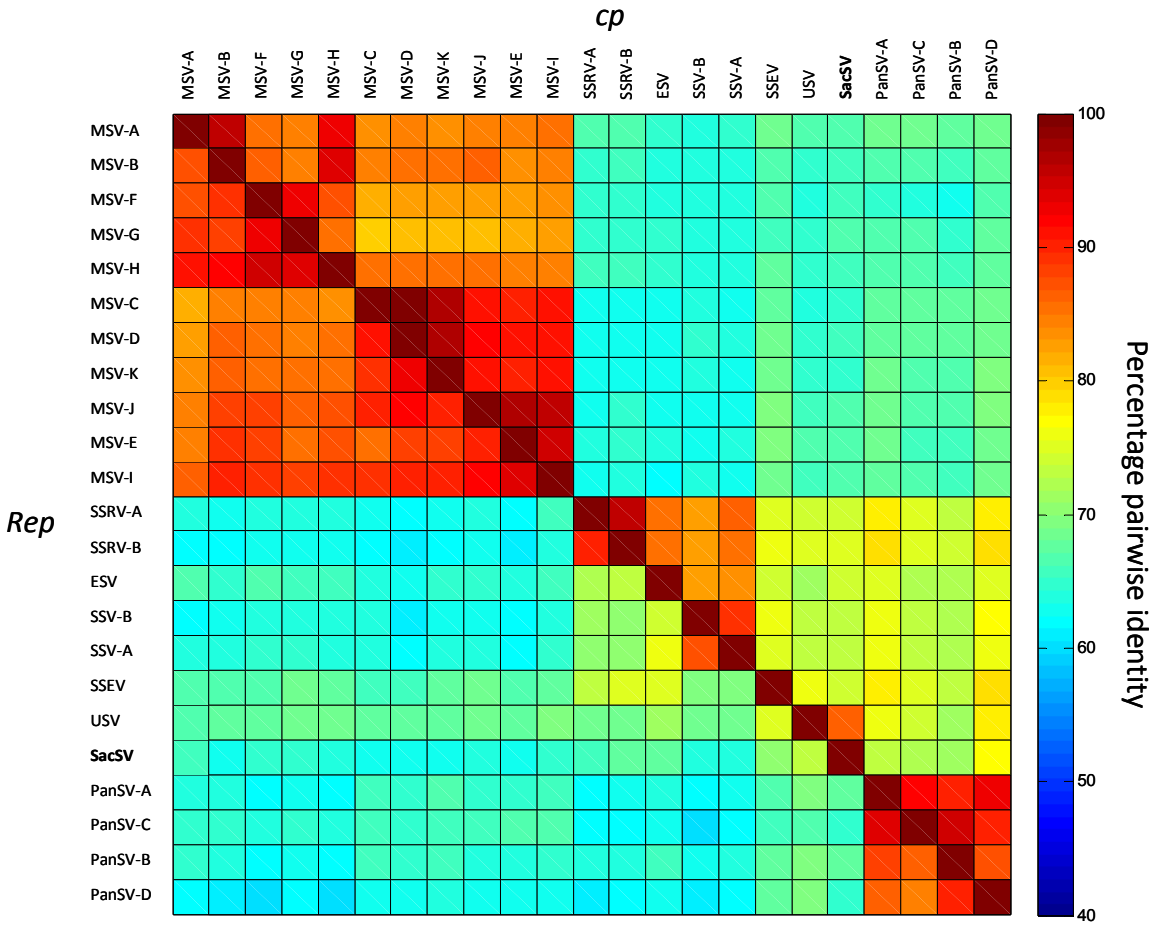


Figure 2.10: A comparison of relatedness between the Rep and CP regions of a selection of mastreviruses. The order of species is based upon the trees above. Plots were constructed based on pairwise distance matrices, showing percent pairwise identity.

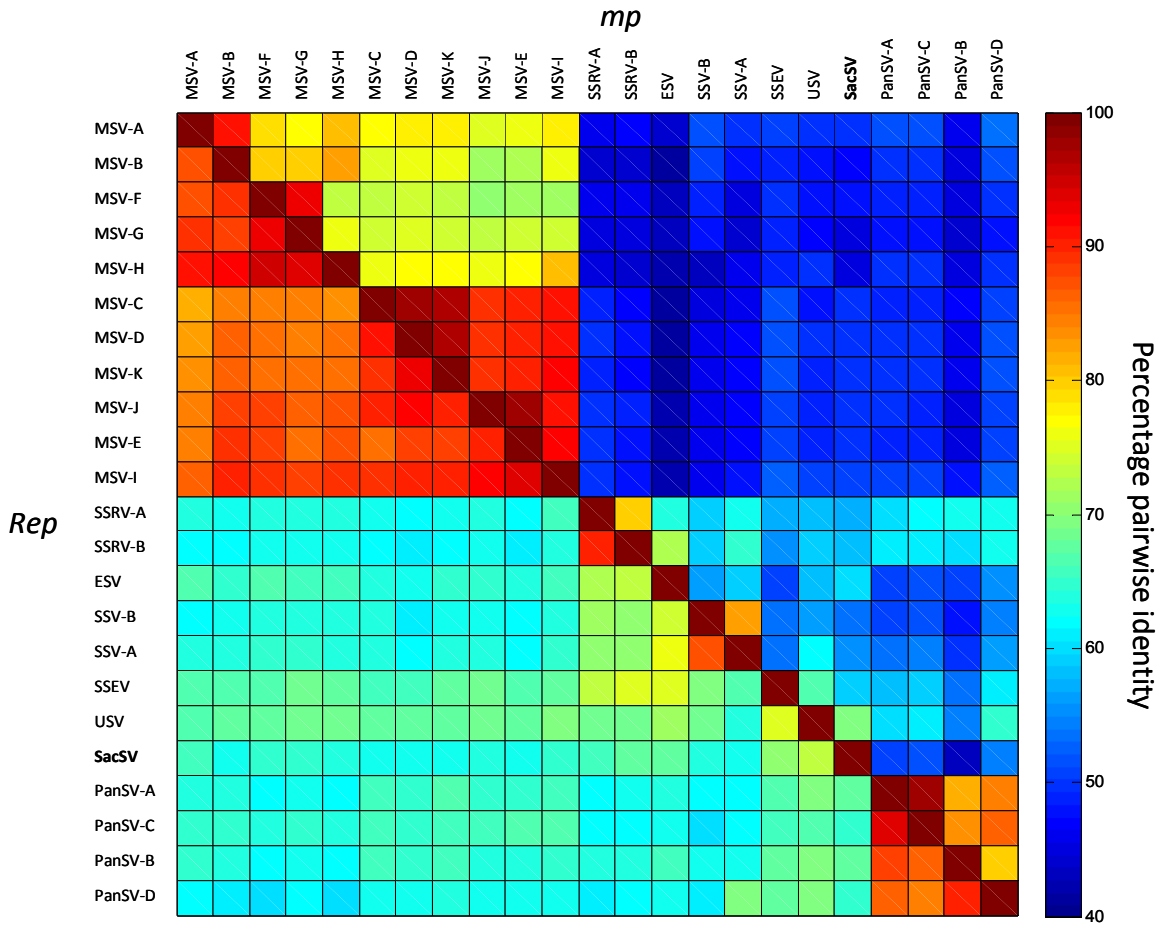


Figure 2.11: A comparison of relatedness between the MP and Rep regions of a selection of mastreviruses. The order of species is based upon the trees above. Plots were constructed based on pairwise distance matrices, showing percent pairwise identity.

2.5 Discussion

In the last couple of years, as a result of the use of sequence non-specific rolling circle amplification with ϕ29 polymerase, significant advances have been made towards both mapping the extent of African mastrevirus diversity and identifying indigenous grasses that are the natural hosts of these viruses (Oluwafemi *et al.,* 2008, Owor *et al.,* 2007b, Pinner *et al.,* 1990 , Shepherd *et al.,* 2008, van der Walt *et al.,* 2009, Varsani *et al.,* 2008a,2008b). Given that we have now discovered yet another new species of African streak virus that infects sugarcane, a worthwhile complement to these efforts could be the wide-scale screening of sugarcane for mastreviruses. Besides such an endeavor possibly yielding evidence of new African streak virus species, it should also help quantify the collective risks of different mastrevirus species emerging as serious constraints on African sugar production.

Notable from this study is the differing levels of relatedness across segments of the mastrevirus genome. SacSV itself shows considerable evidence of this, with relatedness at full genome level being closest to USV, yet being more related to SSEV within the movement protein (Figures 2.5-2.8). Also interesting to note is the much lower levels of relatedness between MSV strains within the coat protein region, particularly of the K, D and C strains. This data corresponds well with known recombination patterns (Varsani *et al*., 2008a). The high relatedness between the A and B coat protein corresponds well to a recombination event present within this region, but not present in the K, D and C strains. This perhaps highlights the ability of certain subsections of the genome to move by recombination, potentially creating different patterns of infectivity. An interesting finding is the higher relatedness between the Rep and CP regions of all species (Figure 2.10). This could possibly be the result of the complex interactions these two regions have with each other and other sections of the genome (Martin *et al*., 2005). MP-Rep and CP-MP show much less relatedness, although differing identities between phylogenetic groupings are still observable (Figures 2.9 and 2.11). MP appears to share the least identity in all paired datasets.

Importantly, it has now been demonstrated that mastreviruses have tremendously high mutation rates (Harkins *et al,*. 2009b, van Antwerpen *et al.,* 2008) and are capable of rapid host adaptation through recombination (van der Walt *et al.,* 2008). The potential threat of emergent African streak virus genotypes damaging sugarcane production on the continent is probably reasonably high and should be taken seriously.

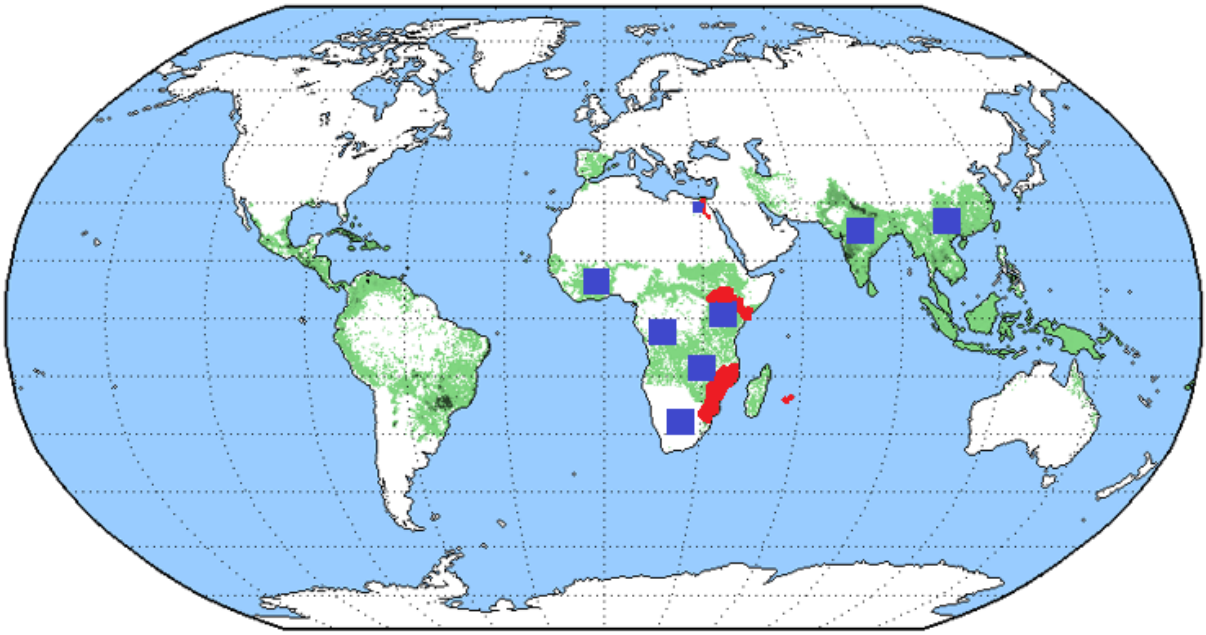


Figure 2.12: Distribution of sugarcane worldwide. Green denotes sugarcane distribution (Monfreda et al., 2008), blue indicates general areas containing insect vectors of mastreviruses, and red denotes regions with SSV infections (Biggare et al., 1999). Note the numerous growers in China and India, both of which have documented cases of other Mastreviruses and their vectors. Sugarcane distribution mapping courtesy of the University of Minnesota Institute on the Environment,

In particular, formal steps should be made towards both monitoring the occurrence of mastreviruses in sugarcane and preventing the movement of infected material both within Africa and between Africa and the rest of the world. Monitoring of the movements of infected sugarcane may in fact be crucial due to the geographical distribution of sugarcane growing countries (Fig 2.12) many of which share similar climates to Africa, and some of which already contain other mastrevirus species and their vector insects (Biggare *et al.,* 1999). The high persistence of mastrevirues in their vectors further aggravates this issue (Ammar *et al.* 2009), making the likelihood of a colonization event

significantly higher. Indeed given the distribution of mastreviruses on Indian ocean islands, is likely that such events have previously occured. This threat is reinforced as the exact factors responsible for the high mobility of some mastreviruses are unknown, and could potentially arise in Sugarcane Infecting Streak Viruses (SISV), which have similar patterns of evolution. The potential for rapid spread of SISVs, and the economic damage this could cause further reinforce the need for thorough monitoring programs to be instigated.

Steps should also be taken to investigate the high sensitivity of sugarcane to infection as this may shed light upon important host-virus interactions or on host defencedefense against mastreviruses. The high variation seen between the numerous species of SISVs (Biggare *et al.,*1999, Hughes *et al.,* 1993*;* van Antwerpen *et al.,* 2008) (Fig 2.5) could indicate that a variety of such mechanisms may be observed and if so may reveal numerous modes of virus attack upon a host. This may prove especially interesting when compared to maize streak virus species, all of which share a high identity and are known to have a similar mode of action. This would require in depth study of the mechanisms of action of each of the SISVs and could shed light upon the types of gene and protein interactions that affect virulence of mastreviruses. This will further our ability to link certain gene types with virulence, and possibly improve our ability to predict subsets of viruses that are likely to be virulent in other host plants. It may also allow us to uncover the reasons behind the seemingly limited modes of action of MSV in maize. This then aids us in preventing virus spread as containment measures based on known risk factors in certain populations of virus can then be implemented.

Genebank accession #

SacSV-[ZA-Emp-T1-2008]: GQ273988

3 Selection Patterns and Modularity in the Mastrevirus Coat and Movement Proteins

3.1 Abstract

Coat protein and movement protein are associated with toxicity and cellular movement in the mastreviruses. Initial studies with smaller datasets identified conserved genomic regions and amino acid motifs within each of these proteins. Recently the number of sequences available has increased significantly, making larger scale conservation and selection studies possible. Studies have linked certain viruses to particular host plants and 3D models of the coat protein of certain mastreviruses have been resolved. Using this information, we aimed to identify conserved regions and regions under selection within species and amongst species infecting certain host plants. We also attempted to map this information to 3D models, to allow easier visualisation and identification of functional regions. Using Clustal W for sequence alignment and the HyPhy toolkit for evolutionary analysis we were able to identify both conserved areas and areas under selection. This data was then successfully modeled to 3D structures using PYMOL. We identified sites under selection across the coat protein and movement protein and discovered several highly conserved genomic regions in the coat protein. This data confirms the significance of the 104 amino acid DNA binding region of coat protein, and gives potential regions for study along the movement protein. The selection detectable across the protein was not significant enough to justify a non-neutral evolutionary model for the mastreviruses.

3.2 Introduction

The mastreviruses are a diverse group of *Geminiviruses* found predominantly in Africa and Indian Ocean islands. These species are grouped as African streak viruses (AfSV), and include Maize streak virus, *Maize streak virus* (MSV; Howell, 1985); *Panicum streak virus* (PanSV; Varsani *et al*., 2008b), *Sugarcane streak virus* (SSV; Hughes *et al.,* 1993); *Sugarcane streak Egypt virus* (SSEV; Biggare *et al.,* 1999*)*, *Sugarcane streak Reunion virus* (SSRV; Shepherd *et al.,* 2008), *Eragrostis streak virus* (ESV; Shepherd *et al.,* 2008) and *Urochloa streak virus* (USV; Oluwafemi *et al.,* 2008*)*. Mastrevirus species also exist in Australia, China, Japan, Europe, Vanuatu and the Middle East, with more recent reports of them being present in Pakistan (Nahid *et al*., 2008). These include *Wheat dwarf virus* (WDV; MacDowell *et al*., 1985), *Chloris striate mosaic virus* (CSMV; Anderson *et al*., 1988), *Digitaria streak virus* (DSV; Donson *et al*., 1987), *Miscanthus streak virus* (MiSV; Chatani *et al*., 1991), *Oat dwarf virus* (ODV) and *Barley dwarf virus* (BDV). The above species all infect monocotyledonous plants. A small group of mastreviruses, including *Bean yellow dwarf virus* (BeYDV; Liu *et al*., 1997a) and *Tobacco yellow dwarf virus* (TbYDV) infect dicotyledonous plants.

Mastrevirus species are classified according to sequence identity thresholds. Strain diversity in AfSVs is classified by a 93% or greater sequence identity to other mastreviruses, and complies with a detectable natural trough in variation around this mark (Varsani *et al.,* 2008a., Martin *et al.,* 2001). The species demarcation for mastreviruses is 75% or lower sequence identity, combined with analysis of differences in key genomic regions (Fauquet *et al*., 2008). This is necessary both because of the widely observed recombination in mastreviruses, and the relatively small differences required to cause vast changes in viral infectivity. This is best illustrated by MSV, in which 91.3% of a sample of 83 sequences showed detectable recombination (Varsani *et al*., 2008a). MSV also illustrates the complexities of gene function in these viruses, as only one of the 11 strains is virulent in maize, even given 93% or greater sequence identity.

The AfSV group is comprised of a number of species, generally sharing around 60%-75% identity. Interestingly, DSV from Vanuatu shows a 67% sequence identity to MSV and is more closely related to AfSVs than other mastreviruses, for example, it shares 47% identity to WDV from Europe (Donson *et al*., 1987). European mastreviruses and those from around the world tend to share around 50% or less identity. WDV shares 47% identity to its most closely related AfSV. A ML phylogenetic relationship of all mastreviruses is provided in Figure 3.1. Host species may also have evolutionary effects on the mastreviruses, both due to shared selection pressures and may have impacts on evolution through recombination in order to rapidly explore sequence space. It has been suggested that the ability of numerous species and strains of mastreviruses to infect a range of plant species may heighten the chances of inter-species recombination (Lefeuvre *et al*., 2007b).

Full sequence identity does not always provide us with sufficient information for accurate analysis. Although full sequence identity allows grouping by species or strain, the open reading frames (ORFs) can reveal significant information. Relationships within ORFs are crucial to understanding the context of particular genetic material, particularly in highly recombinant genomes. Mastreviruses can be an extreme case of this, as they possess only four ORFs and a genome of only 2.7-3kb. This makes every ORF highly significant. Relatedness in these regions can give a much better indication of the particular ability of each species in areas such as virulence and host specificity. Owing to the recombinant nature of mastreviruses, relatedness in each section of the genome can potentially hint at the evolutionary origins of composite genomes.

The coat protein and movement protein are important in the ability of mastreviruses to systematically infect a host, and also function in intra-cell movement. They are known to interact with each other to perform these functions as well as functioning best when related cassettes are inherited together (Liu *et al*., 1999, 2001). These two proteins are therefore crucial to understanding the ability of mastreviruses to infect specific host species. Analysis of sequence conservation and selection within these regions could allow better identification of important amino acid sequences within the proteins.

Due to the evolutionary factors affecting different regions of the genomes, relatedness within different sections of the genome is not always similar to overall genetic relatedness. As CP and MP form a gene cassette the relationship between them is important to any study of their evolution. Figures 3.2 and 3.3 show the relationships between the mastrevirus species at the protein level. No detailed study of the selection pressures operating upon the CP and MP have yet been carried out. Here we performed a comprehensive analysis of the selection pressures affecting the coat and movement protein. We also identify key regions using sequence conservation data and 3D modeling, in order to obtain a better understanding of significant parts of these proteins.

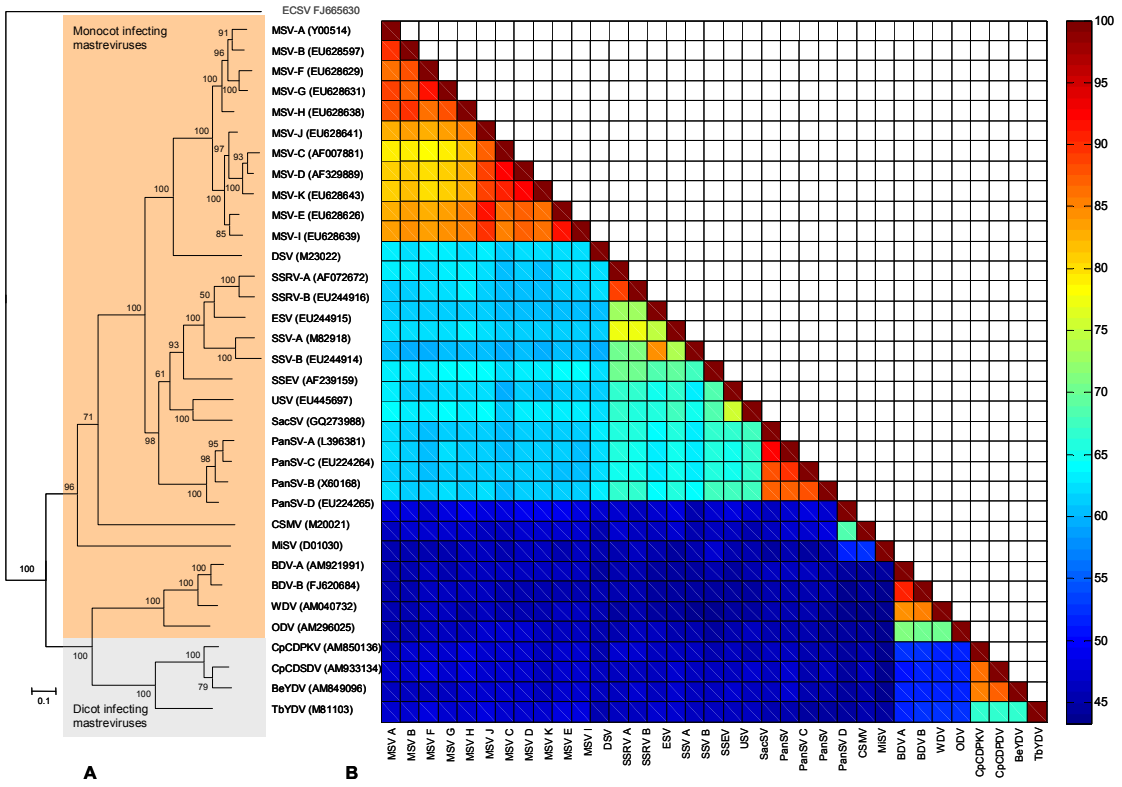


Figure 3.1: A. Maximum likelihood (ML) tree based on an alignment of complete genome nucleotide sequences depicting the evolutionary relationships between masterevirus species and strains. The ML tree was constructed using PHYML (Guidion and Gascuel, 2003) with HKY chosen as the best fit model by ModelTest (Posada 2006).. The numbers associated with tree branches are indicative of the percentage of 1000 full maximum likelihood bootstrap replicates supporting the existence of the branches. **B**. Two-dimensional graphical representation of pairwise genome-wide nucleotide sequence similarities (calculated with pairwise deletion of gaps; scale represents percentage identity) between representative mastrevirus species and strains.

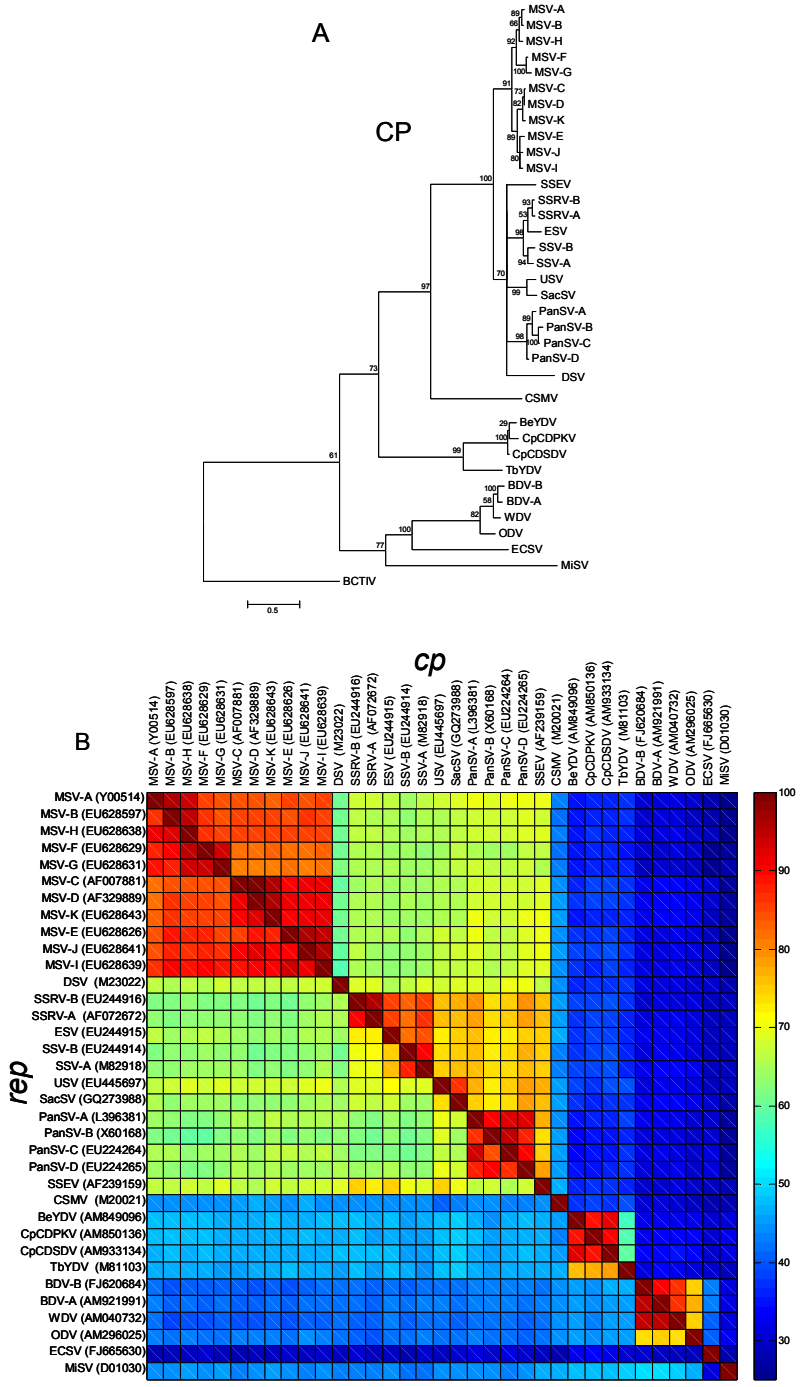


Figure 3.2: A Maximum likelihood (ML) phylogenetic relationships based upon alignments of the predicted amino acid sequence of CP. The ML trees were constructed using PHYML (Guidion and Gascuel 2003) (best fit Rep and CP models = LG as determined by PROTEST (Abascal *et al*., 2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21: 2104-2105). Numbers associated with tree branches are indicative of the percentage of 1000 full maximum likelihood bootstrap replicates supporting the existence of the branches. **B** Two-dimensional graphical representation of pairwise amino acid sequence identities (calculated with pairwise deletion of gaps; scale represents percentage identity) of the predicted Rep and CP of representative mastrevirus species and strains.

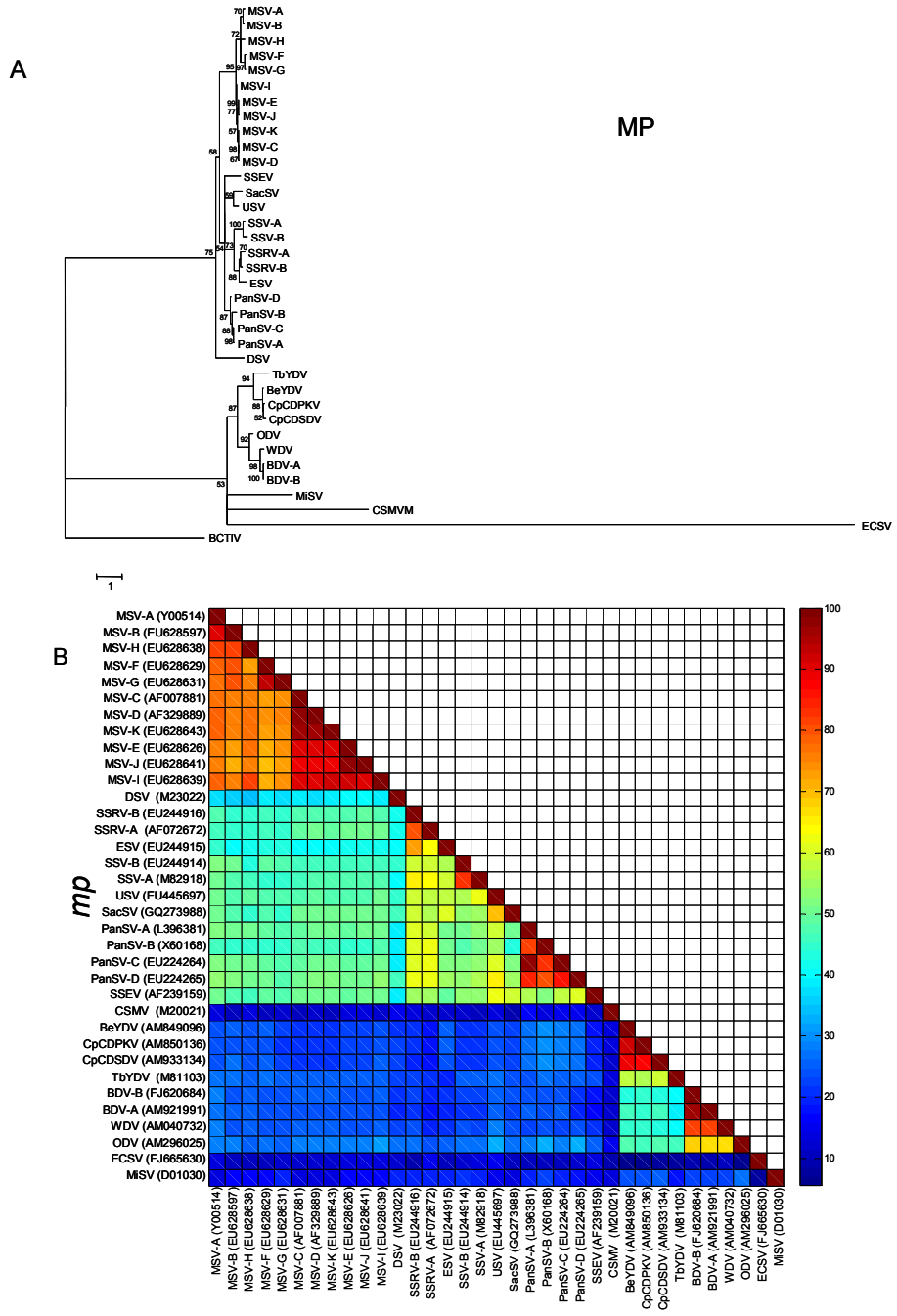


Figure 3.3: A Maximum likelihood (ML) phylogenetic tree depicting relationships between MP amino acid sequences. The ML trees were constructed using PHYML (Guidion and Gascuel 2003); with best model = LG as determined by PROTEST (Abascal *et al*., 2005). Numbers associated with tree branches are indicative of the percentage of 1000 full maximum likelihood bootstrap replicates supporting the existence of the branches. **B** Two-dimensional graphical representation of pairwise amino acid sequence identities (calculated with pairwise deletion of gaps; scale represents percentage identity) of representative mastrevirus species and strains.

3.3 Methods

Sequence Analysis

Sequences were obtained from the Genbank database in the case of DSV, MiSV, BeYD and WDV. Other unpublished sequences (MSV A-K, PanSV, SSRV, SSV) were obtained from the MSV (Martin, Shepherd and Varsani) research group.

Sequences were partitioned 'by species' and then aligned using Clustal W (Thompson *et al.,* 2004), followed by manual alignment in Mega 4.0 (Tamura *et al.,* 2007). Open reading frames (ORFs) were identified by homology to known sequences available on the BLAST database. Movement protein and coat protein coding regions were identified and conserved sequences were recorded. Sequences were then reorganised into two datasets, one arranged by host species, the other by either species or strain depending on whether 3 or more sequences were available (3 being the minimum number allowing for analysis with the HYPHY package).

Codon Selection Analysis

Datasets were realigned at the protein level and reconverted to a nucleotide FASTA file. Stop codons were then removed to conform with HyPhy specifications. Datasets possessing 3 or more sequences were then uploaded to the Datamonkey web server which implements the HyPhy bioinformatics package (Pond *et al*., 2005a, Pond & Frost., 2005b). Each data set was run through a model testing procedure in order to select the appropriate evolutionary model. GARD analysis was then run where possible in order to account for the effects of recombination. Three selection detection algorithms, single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL) and internal fixed effects likelihood (IFEL) were then used to detect positive and purifying selection (Pond & Frost, 2005c, Pond *et al*., 2006).

3D Models

A 3D model of MSV was kindly provided by Mr Kyle Dent (University of Cape Town, South Africa; Varsani research group) which was modeled on the 3D structure of satellite tobacco mosaic virus with fitting into a cryo-EM model of MSV This model was used for 3D modeling, all using MODELLER (Eswar *et al*., 2006) toolkit implemented through the Easy Modeller GUI for this study. PyMOL (Delano, unpublished) was then used to display a space filling model and cartoon form of the proteins. Conserved sites and sites under selection were modelled onto these proteins in an attempt to illustrate links between likely functional and non-functional regions under selection.

3.4 Results

3.4.1 Sequence Conservation of Mastrevirus Coat Protein

Conserved sequences are an important indicator for key functional or structural regions at the protein level. To identify these regions, conserved amino acids were mapped to 3D models (capsomer of 3 subunits) across various ranges of species and groups. Firstly, conserved sites over the entire dataset were mapped (Fig 3.4). Conserved sites were predominantly in random loops. However, a large number of conserved regions were either basic residues or regions flanking basic residues. Only 41 amino acids were conserved across all sampled mastreviruses.

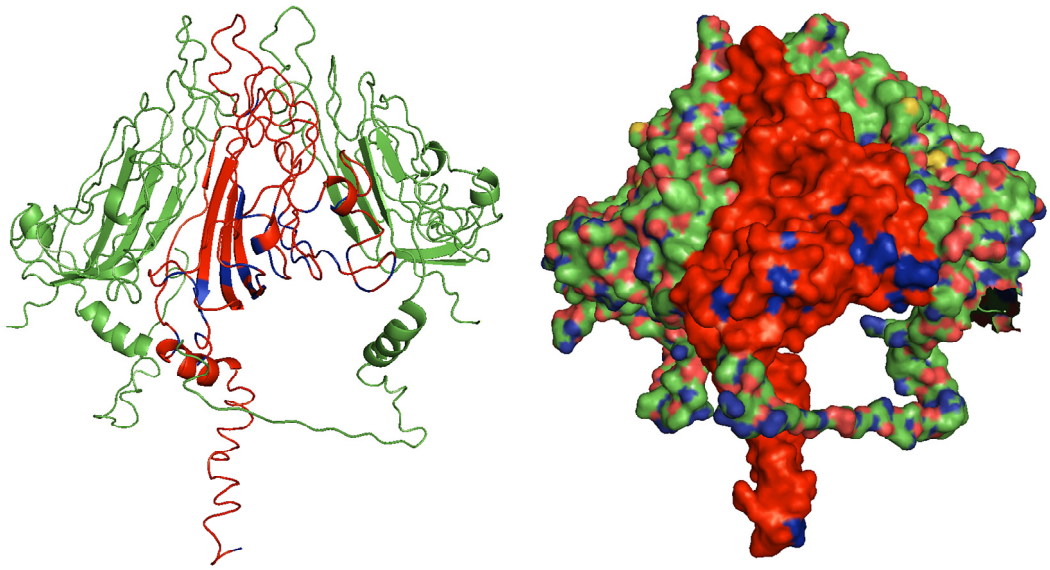
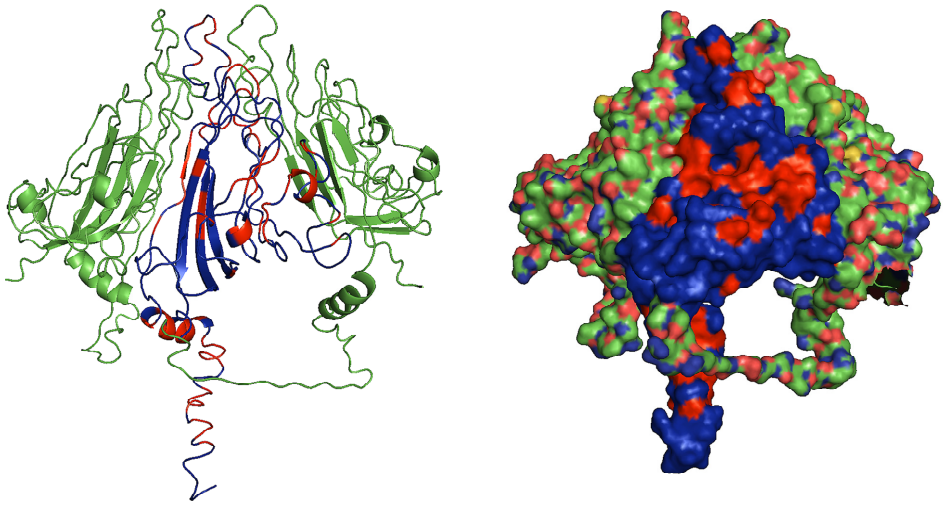


Figure 3.4: Amino acid conservation within the coat protein of all mastreviruses shown in cartoon form and using a space filling model. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1.

Conserved sites between different groups were then examined. The dataset was split into the MSV A-K strains and the sugarcane infecting streak virus group based on branch points on the CP phylogenetic tree (Fig 3.3). Within the MSV-A-K, group conservation was generally found around the interface between different subunits of the capsomer. Highly variable sites on the surface are notable. This conforms with surface regions being more likely to interact with host factors. The Sugarcane infecting streak virus group showed more diversity as expected, however, there was also a slight tendency for conserved regions to be along the interface of subunits.

A



B

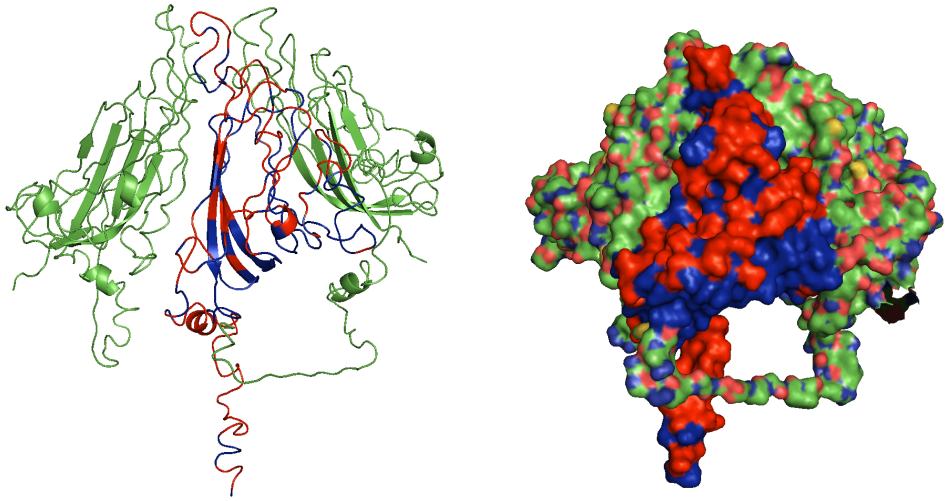


Figure 3.5: A: MSV A-K conservation. **B** SISV conservation. Amino acid conservation mapped to a cartoon structure and a space filling model. Red denotes variable sites. Blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

The dataset was then broken down further into groupings 'by species' or strain, depending on the number of available sequences. The MSV-A dataset was the largest and provided the highest resolution information in this study. A large number of variable sites were clustered around the N-terminal region of the protein (Figure 3.6). Notable also were variable sites within the alpha helices and beta sheets of the protein. Nearly all basic residues were conserved throughout the dataset. Variable regions were predominantly amongst non-polar amino acids.

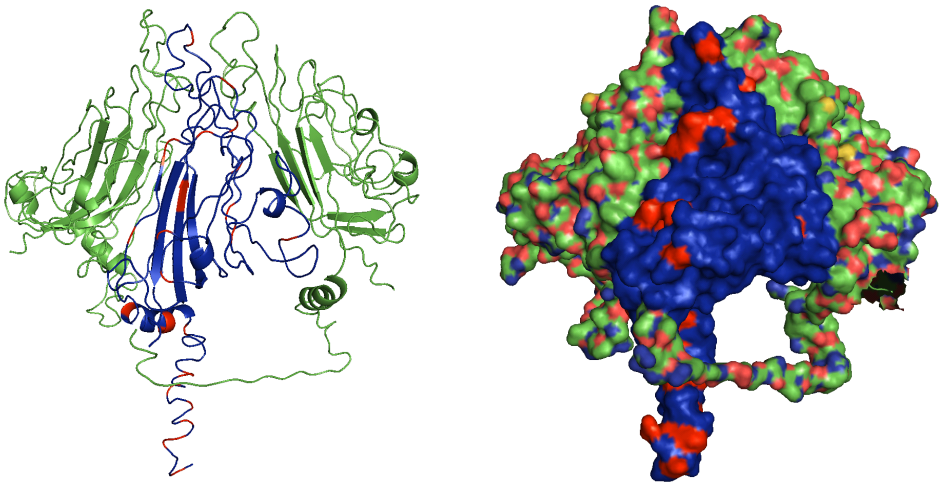


Figure 3.6: Conservation amongst MSV-A, shown in cartoon form and as a space filling model. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

The Panicum streak virus data set shows more variation then that of the MSV-A set, due to the inclusion of all PanSV strains. Variable regions are in similar areas to that of MSV A-K. A band of variable sites around the middle section of the protein is observable in the A-K, PanSV and to some extent in the SSRV and SSV datasets (Figure 3.7). The topmost region of the loop structure contains variable regions in all of these datasets. These variable regions are most often composed of hydrophobic or polar but uncharged residues, thus the functional relevance of this region may be small. Another region of high variability is the N-terminal residue. This is again visible in all of the above datasets, and is potentially more functionally relevant. Variable sites in this region often contain basic residues such Lysine and Arginine, and variation in regions flanking these is common. The USV dataset is entirely conserved at the protein level (Figure 3.8) and this is as a result of the low diversity observed in the USV isolates.

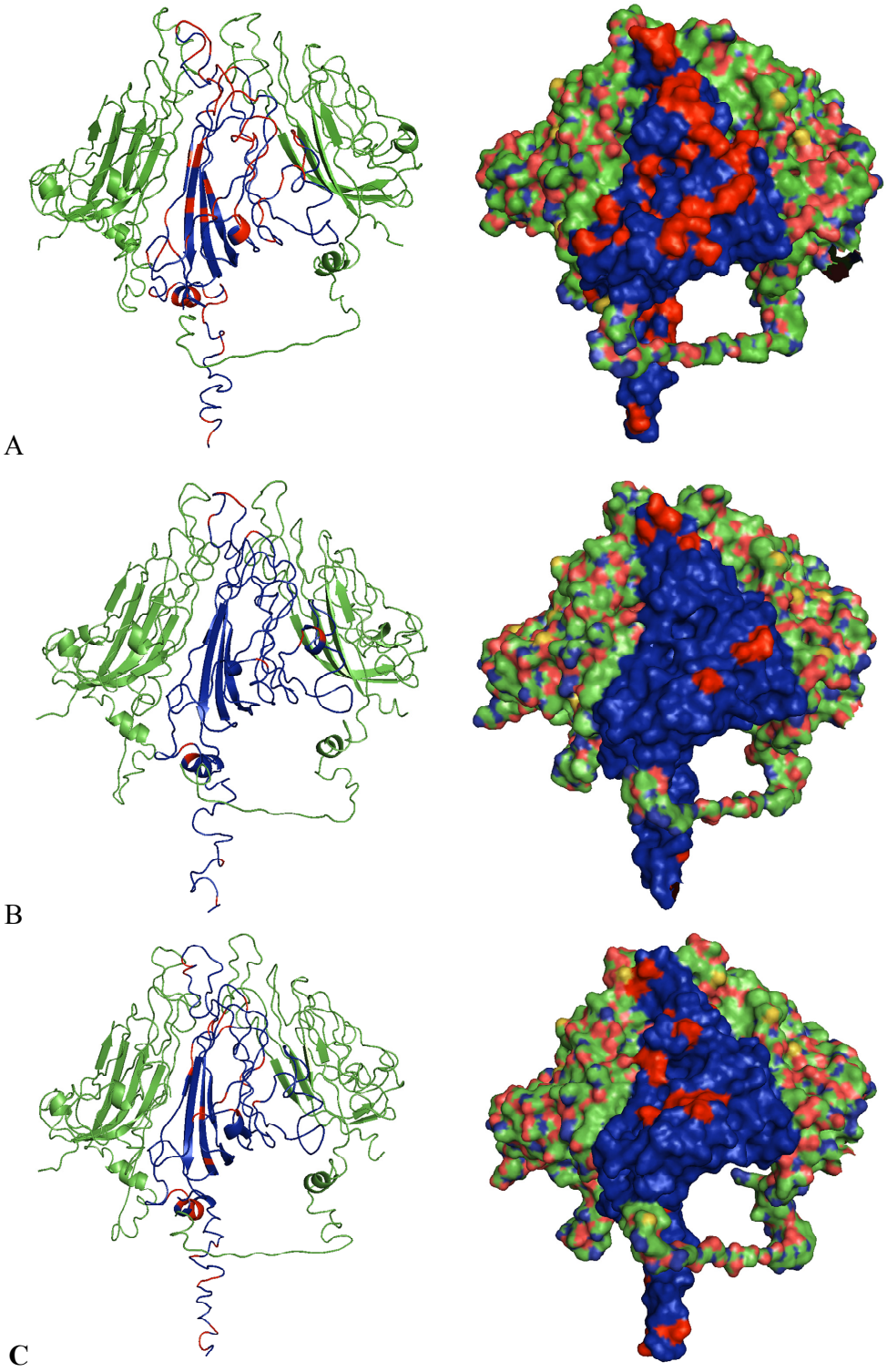


Figure 3.7: A: Conservation within the Panicum streak viruses. B: Conservation within the SSRVs. C: Conservation within the SSVs. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

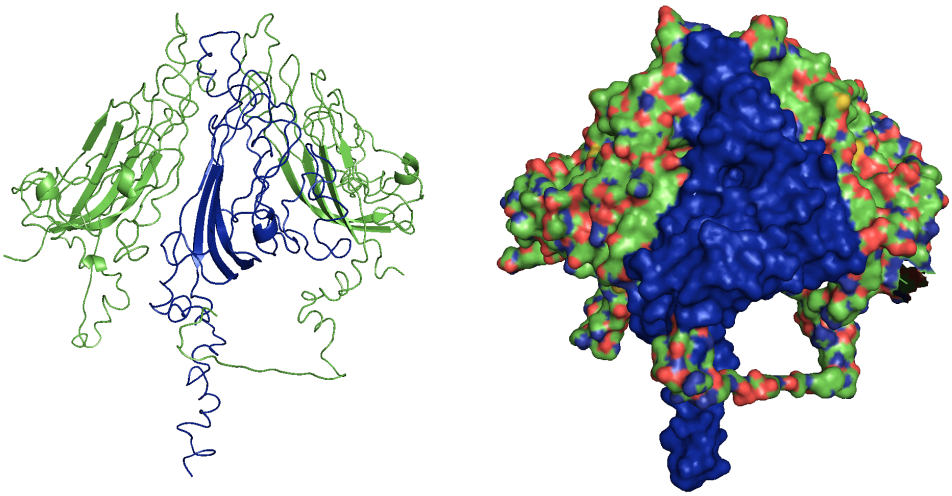


Figure 3.8: USV conservation. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

Over all datasets, two notable regions were observed. A group of basic amino acids **KRKR** in the N-terminal region of the protein was highly conserved. This coincides with the beginning of a known nuclear localisation signal (NLS) and may thus have an important effect on this activity. A second region **LQIQ** was conserved in the majority of datasets, and is positioned approximately twenty resides from the end of the NLS. This region is also within the DNA binding domain. The DNA binding domain is situated along the left side of each subunit. Variation along this side of each unit may therefore reflect either inter-unit interactions or interactions to do with DNA binding. SSRV in particular has a highly conserved binding region. The inside face of the barrel structures surface is noticeably less variable then the outside face. Within the central regions of the protein a sequence **WL**x**YD** was conserved in all mastreviruses, and was positioned on the inside face of the middle region. The C-terminal region **RLYFKSVGNQ** is conserved within all SSV and MSV species, but not over all mastreviruses, differing slightly in BeYD and DSV. WDV's C-terminal region varies significantly from this, with **ACYFKAIGIQ** being the end sequence in all WDV samples used, with the exception of barley infecting strains. However a general sequence xx**YFK**xx**G**x**Q** is still observable even accounting for WDV (Figure 3.9).

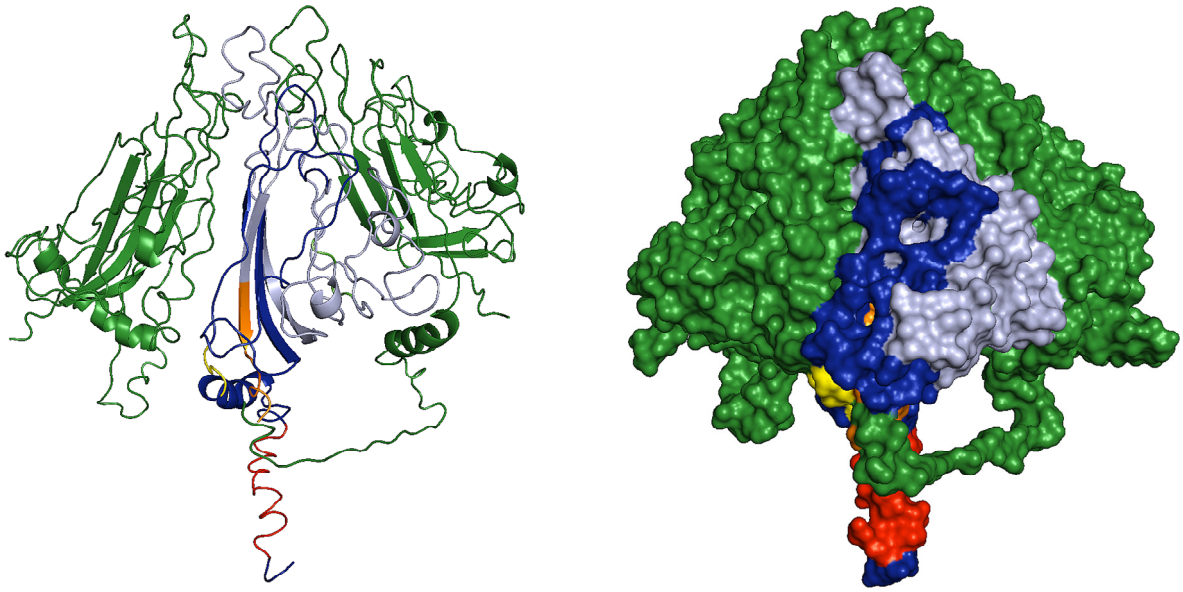


Figure 3.9: A 3D model of MSV-A displaying key sites. Colours denote various conserved sites. Forest Green: Flanking capsomers, drawn here to show interface regions more clearly. Grey: Non-conserved regions on the capsomer. Blue: 104 amino acid DNA binding region. Orange: The conserved RLYFKSVGNQ sequence. Yellow: Conserved LQIQ sequence. Red: Location of NLS. Light Green: WLxYD conserved sequence.

3.4.2 Sequence Conservation of Mastrevirus Movement Protein

The movement protein was highly conserved within species, but had few sites conserved over all mastreviruses. Central regions near or within the transmembrane domain tended to be more conserved than end regions. An **LKAR** site was conserved across the sugarcane infecting streak viruses. A sequence **TEE** near the C-terminal of the potential transmembrane domain was conserved. In the 'by host' datasets central regions were again more conserved. Both basic and acidic charged residues were also highly conserved, with the exception of the C-terminal regions, which contained numerous non-conserved basic residues in all by host datasets.

3.4.3 MSV-Kom and MSV-Set Comparison

MSV-Kom and MSV-Set have been used numerous times to compare virulent and non virulent strains of the maize streak virus in maize. MSV-Kom is a maize-virulent MSV-A isolate, Set is considerably less virulent in maize and is an MSV-C isolate (Martin *et al*., 2001). A comparison of conservation patterns throughout these two viruses is appropriate to examine what may be key sites affecting infectivity. 39 amino acids differed between

the two CPs. Of these 20 were within the first 50 amino acids. Six were within the NLS region. within the NLS regions a highly basic **KKK** region was flanked by a Proline residue on each side in MSV-Kom. MSV-Kom had a slight tendency to use Arginine instead of lysine. In total 25 of the variable sites were within the DNA binding region. Of the remainder two distinct clusters were observed, with 6 sites between positions 120-139 and a further 6 between 160 and 186. A majority of variable sites were again on the outer surface of the barrel structure, or within the N-terminal arm.

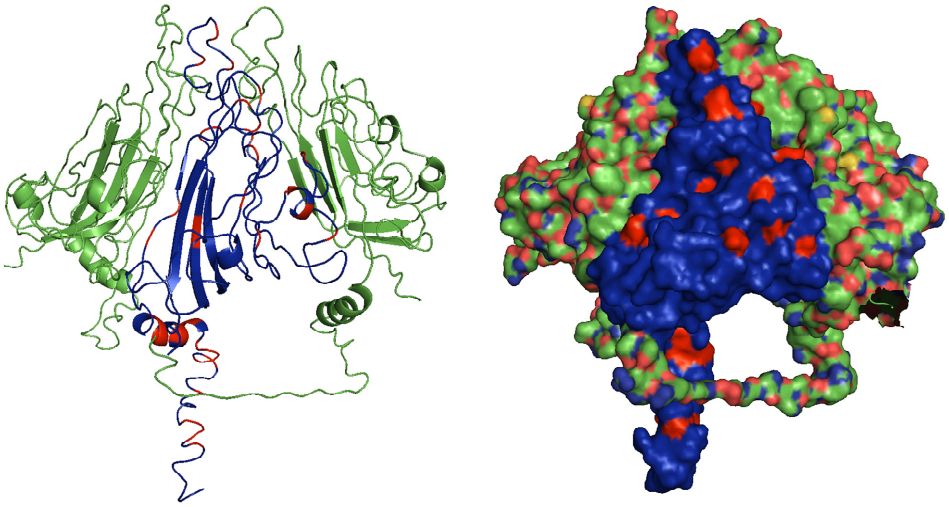


Figure 3.10: MSV-Kom and Set conservation analysis. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

As MP and CP are thought to interact, a conservation study of MP was appropriate for this virulence comparison. 23 of 102 amino acids were variable between the two viruses. Of these 11 clustered between residues 72-92, and 8 between residues 5-22. MSV Kom's movement protein also had a higher number of polar residues, and fewer acidic residues then MSV Set

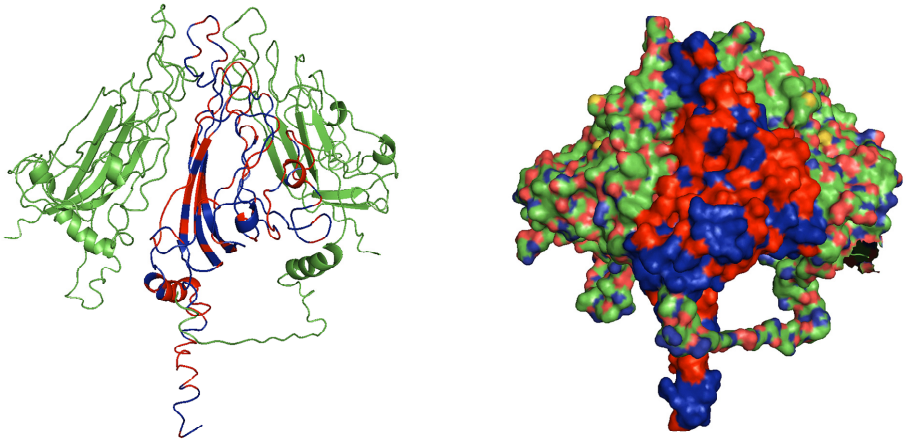
3.4.4 Conservation by Host species

In order to determine the influence of host factors on mastreviruses the conservation patterns amongst viruses that infect specific hosts were examined. Sequences were subdivided according to the findings of Varsani *et al* (2008a). Digitaria infecting viruses were isolates of DSV and MSV A4-B3, E-G and I. Significantly more variation exists in this set than in the MSV-A -K dataset, however, regions around the binding site and the edge facing the binding site are almost as highly conserved (Fig 3.11). Variation in amino acids along the interface between subunits is significantly higher.

The panicum dataset includes the isolates from PanSVs and MSV-G. Variation is again higher, with a similar tendency for the mid-region of the protein to be more variable. In this case however sites along the capsomer interface (left side in the diagram) of the subunit remains reasonably conserved, potentially indicating a more conserved DNA binding region amongst this set.

The setaria dataset consists of isolates of MSV B3, E, H and K. The sugarcane dataset consists of isolates of MSV-A4, SSRV, SSV and SacSV. The Urochloa dataset consists of isolates of MSV A1, B1, B3, F and USV. All datasets show a tendency for increased variability in the central regions when compared with mastrevirus species datasets. The setaria set retains many conserved sites around the DNA binding region/subunit interface whilst sugarcane and urochloa datasets do not. The urochloa dataset shows a much increased variability in the C-terminal regions, although these regions are still less variable than central regions of the protein.

A



B

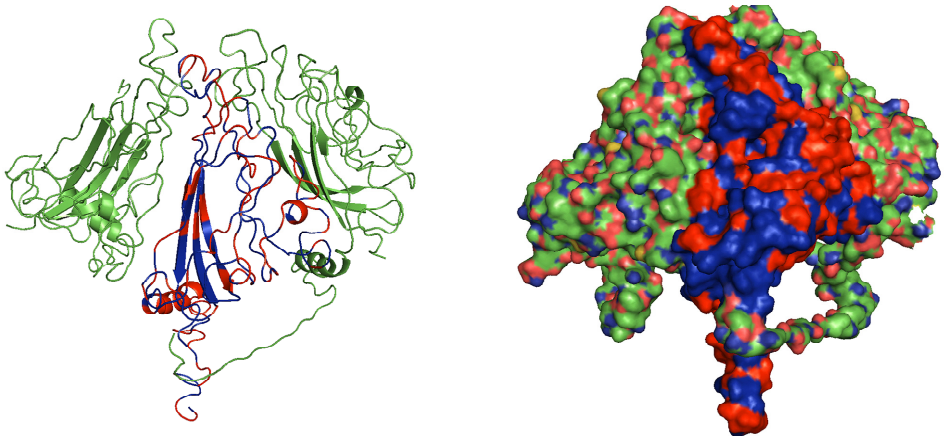
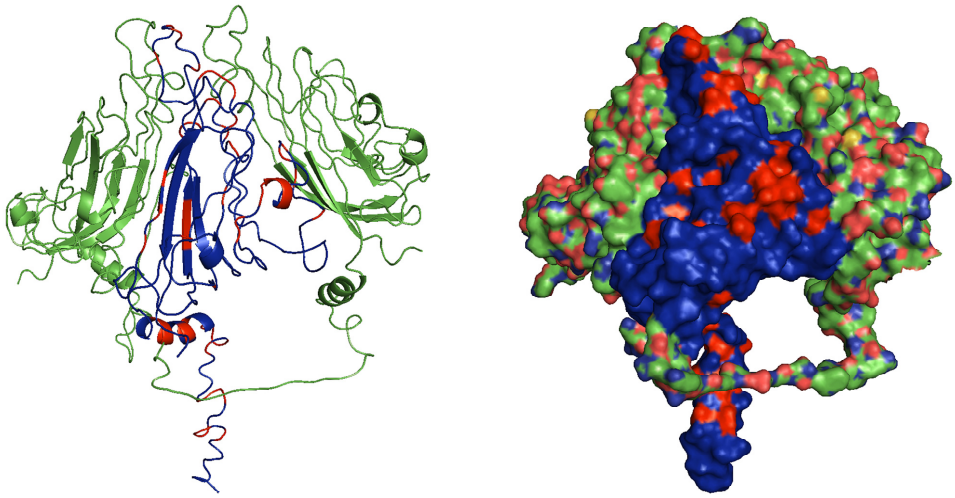


Figure 3.11: A: Digitaria conservation by host. **B**: Panicum conservation by host. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

A



B

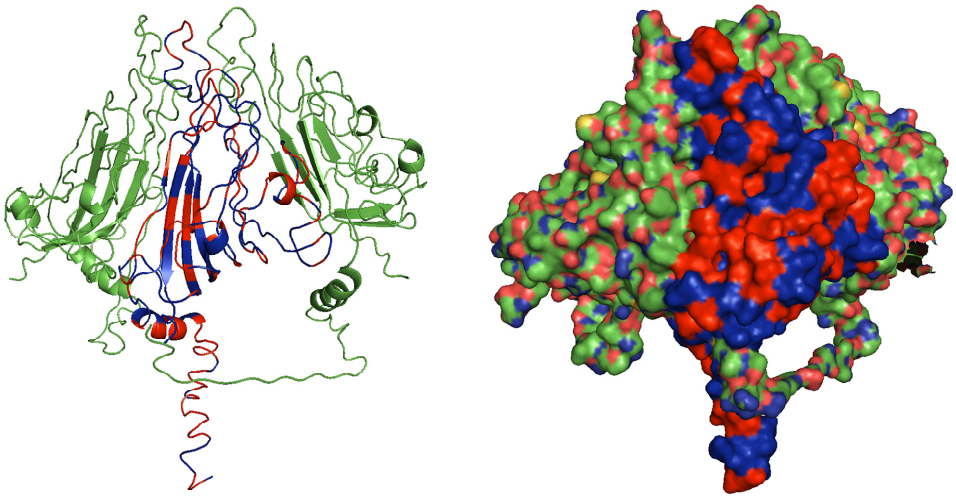


Figure 3.12: A: Setaria Conservation by host **B**: Sugarcane conservation by host. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

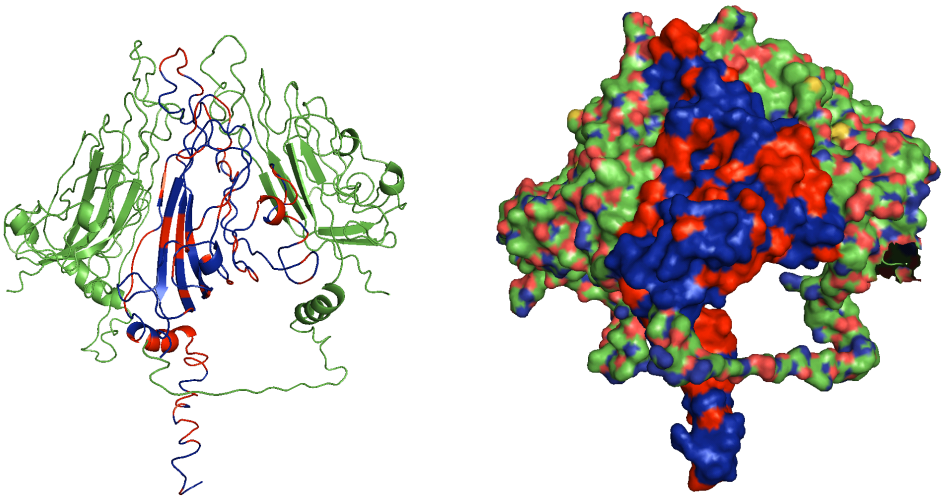


Figure 3.13: Urochloa conservation by host. Red denotes variable sites, blue denotes conserved sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for full alignments.

All datasets again show the highly conserved sequences **KRKR**, **LQIQ**/**A** and C terminal regions. Basic sites seem less conserved in these datasets than within species, however much of the variation that exists is still within hydrophobic amino acids. The existence of highly variable regions across the protein middle surface region in all sets could indicate that this region is less relevant for specific host virus interactions. Combined with the high conservation apparent in certain sites may indicate that key sites are responsible for general infective ability, with variable sites 'tweaking' other factors such as virulence.

3.4.5 Selection

In order to understand which regions of each protein were of evolutionary importance, selection analysis was carried out. SLAC, FEL and IFEL analysis was performed on the datasets in order to give a comprehensive summary. FEL gave the most indications of selection, followed by IFEL and then SLAC, which was as expected due to SLAC being a highly conservative indicator of selection.

Coat Protein

Selection was detected over a number of sites by all three mechanisms. Detectable selection was identified across the length of the coat and movement proteins.

As expected the majority of detectable selection was negative or purifying selection. The MSV-A and PanSV datasets showed strong negative selection along a significant portion of the coat protein (Figure 3.15). The number of sites under detectable selection within the MSV-A -K species was significantly lower, with the exception of MSV B1 (Figure 3.14). USV appears to be under near neutral selection with only one detectable site under negative selection. Positive selection was detectable in the SSV, SSRV, PanSV, MSV-A and MSV-B3 datasets.

The conserved **LQIQ** site and sections of the **KRKR** region were under detectable selection in several datasets (around codon 101). Charged sites, basic amino acids and regions flanking these were frequently subject to negative selection, particularly in the DNA binding region. A concentration of negatively selected sites was detected in both the MSV-A and PanSV datasets around condon 170-179. Several regions appear to be under no detectable selection in all datasets. The C-terminal end of the protein, and areas around codons 51,151 and 201 have relatively fewer selected sites than the remainder of the protein. Positive selection is generally more prevalent in the DNA binding region and in MSV-A positive selection was detectable at two sites in the NLS.

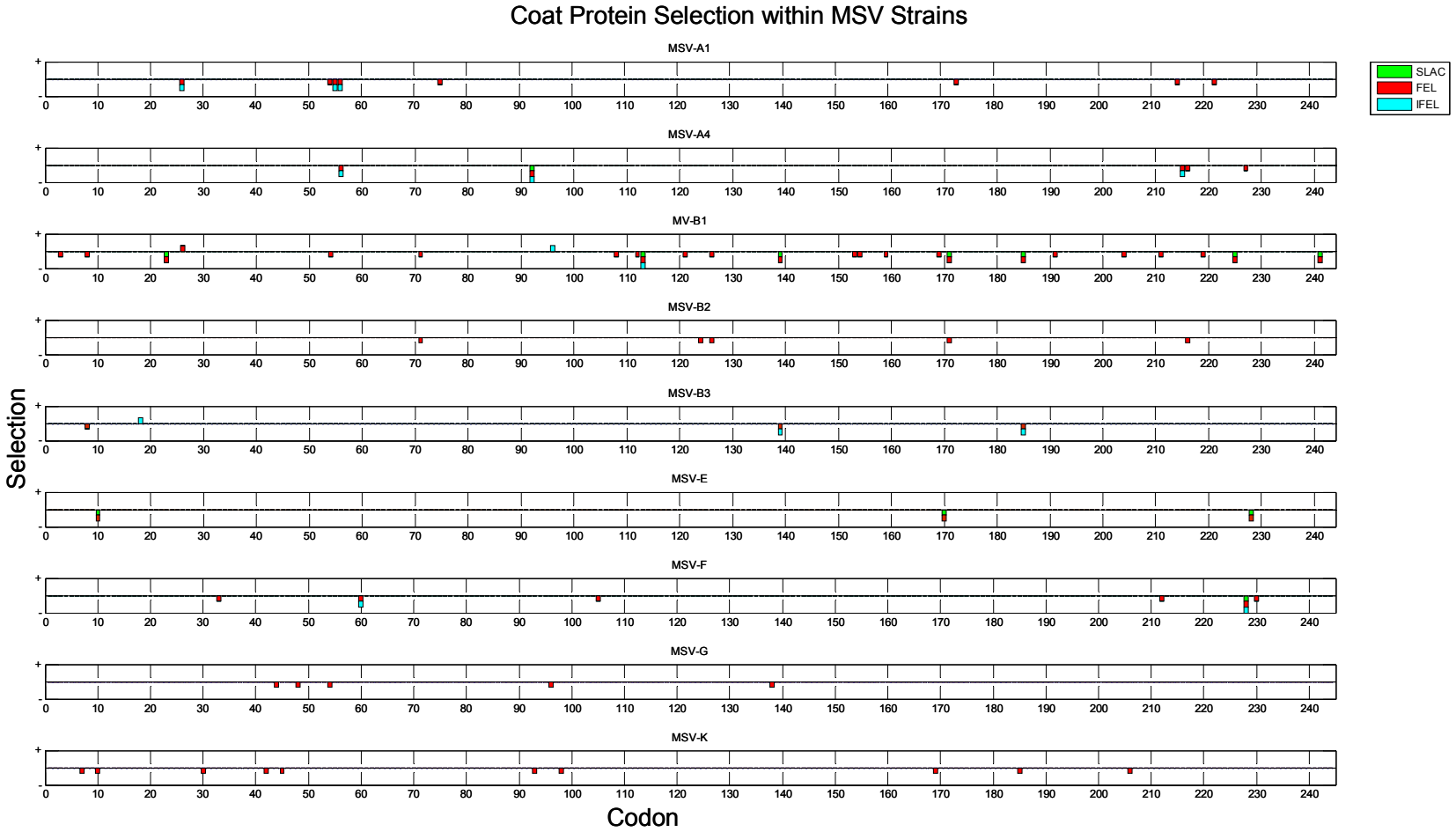


Figure 3.14: Codon selection analysis on MSV strains A-K. Sites under selection detected by SLAC, FEL and IFEL are presented. Selection is denoted by coloured boxes. Full alignments can be found in the Supplementary material.

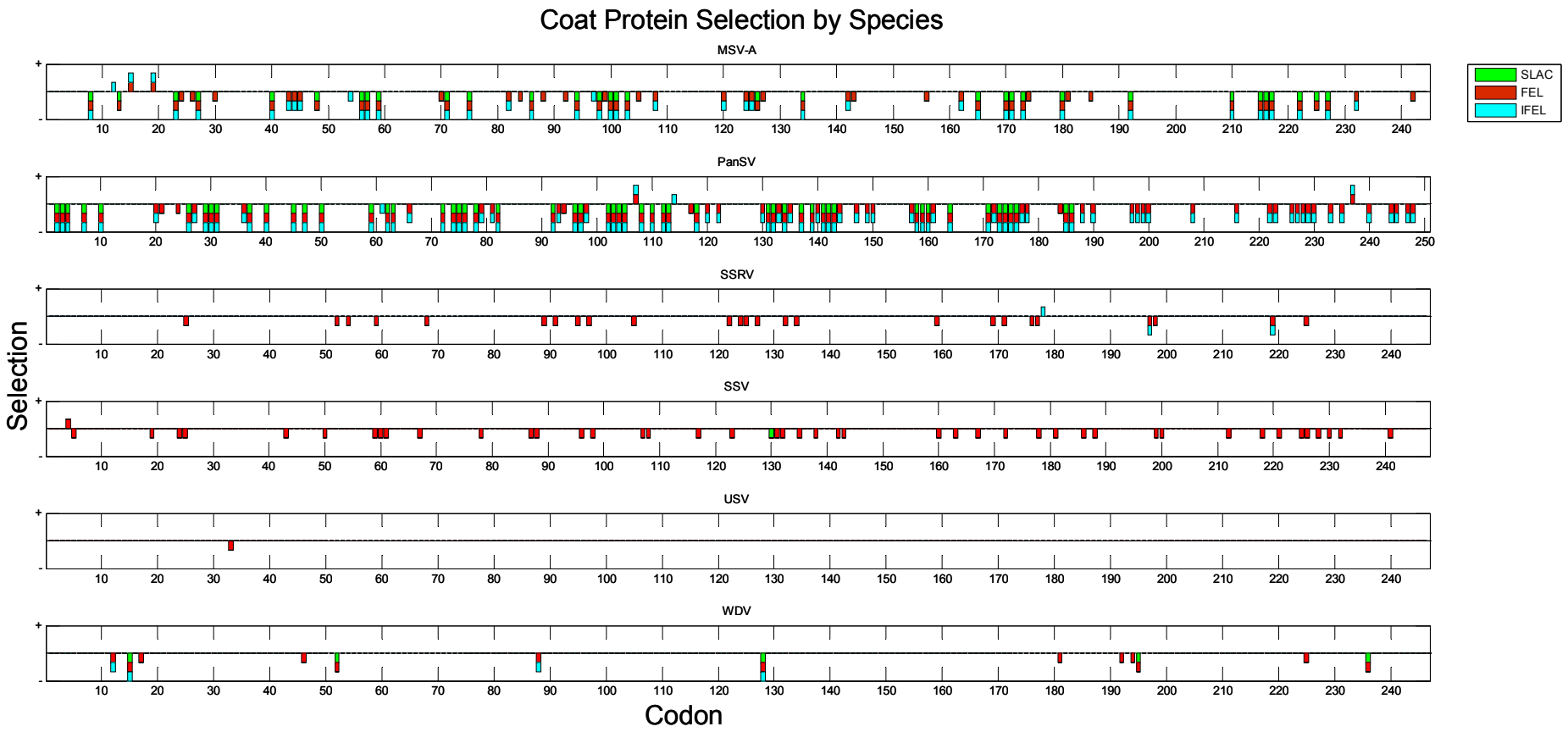


Figure 3.15: Selection analysis on the Coat Protein by Species: Sites under selection as detected by SLAC, FEL and IFEL are presented here. Detection is indicated by coloured boxes, which may be stacked if multiple algorithms detected selection. Green denotes SLAC, red FEL and blue IFEL. An example of the coat proteins structure is presented above to indicate relevant areas. See the Supplementary material for full alignments.

Movement protein

Selection analysis was also run on the movement protein. Detectable selection was spread across the protein, however regions of dense negative selection were detectable in the PanSV and SSRV datasets around codons 101 and 57 respectively. The detectable selection was again predominantly negative, although significantly fewer sites were detected. Selection within MSV strains and USV was very low. Therefore the majority of sites were selectively neutral (Fig 3.16, 3.17). Considerable selection was detectable in the transmembrane domain of the protein, currently the only known functional region. The movement protein of PanSV has noticeably more negative and positive selection than that of other species. Positively selected sites in the movement protein tended to be in C-terminal regions.

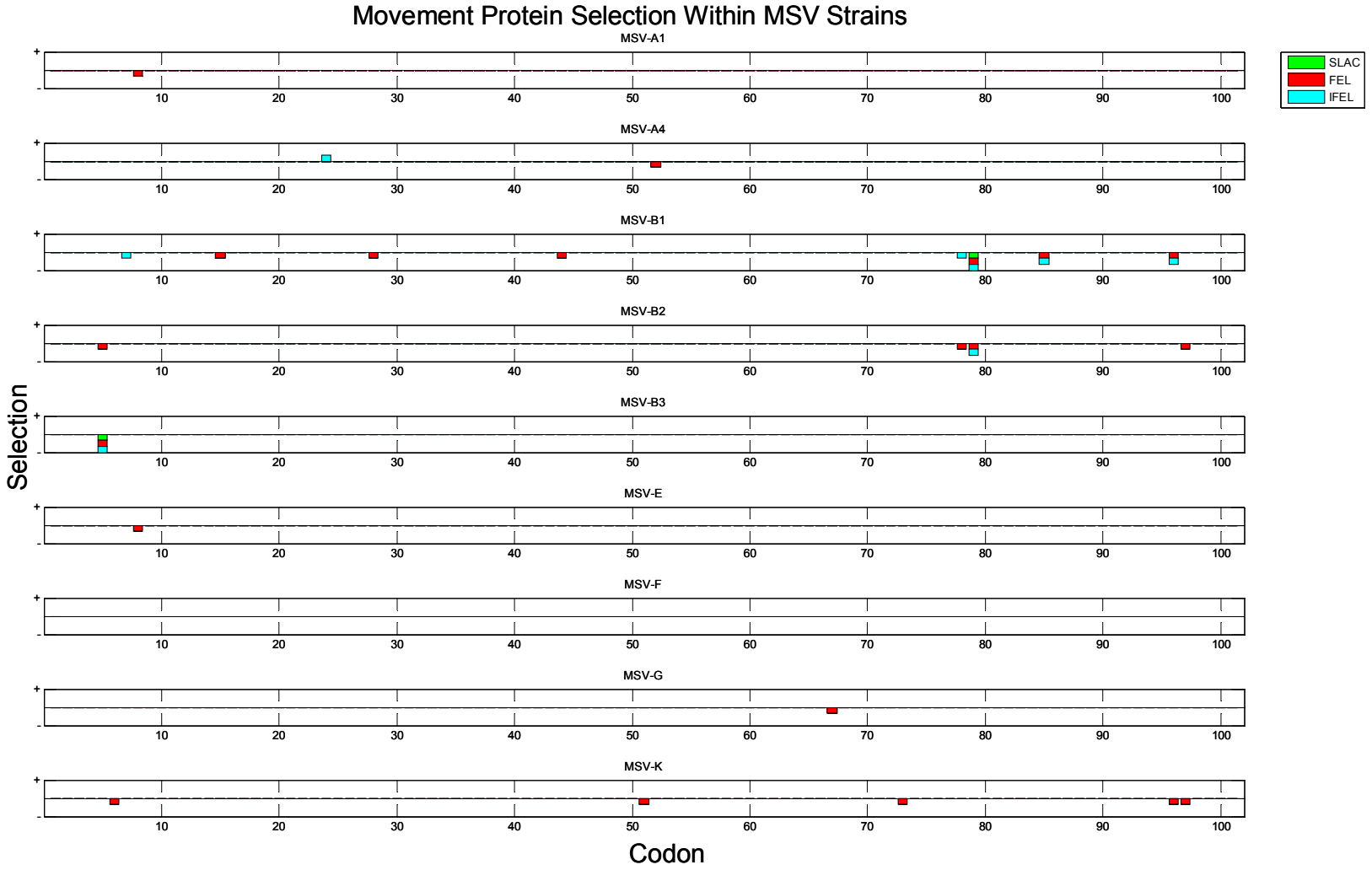


Figure 3.16: Selection analysis on the movement protein of the MSV strains A-K. Sites under selection are indicated by coloured boxes. For full alignments see the Supplementary material.

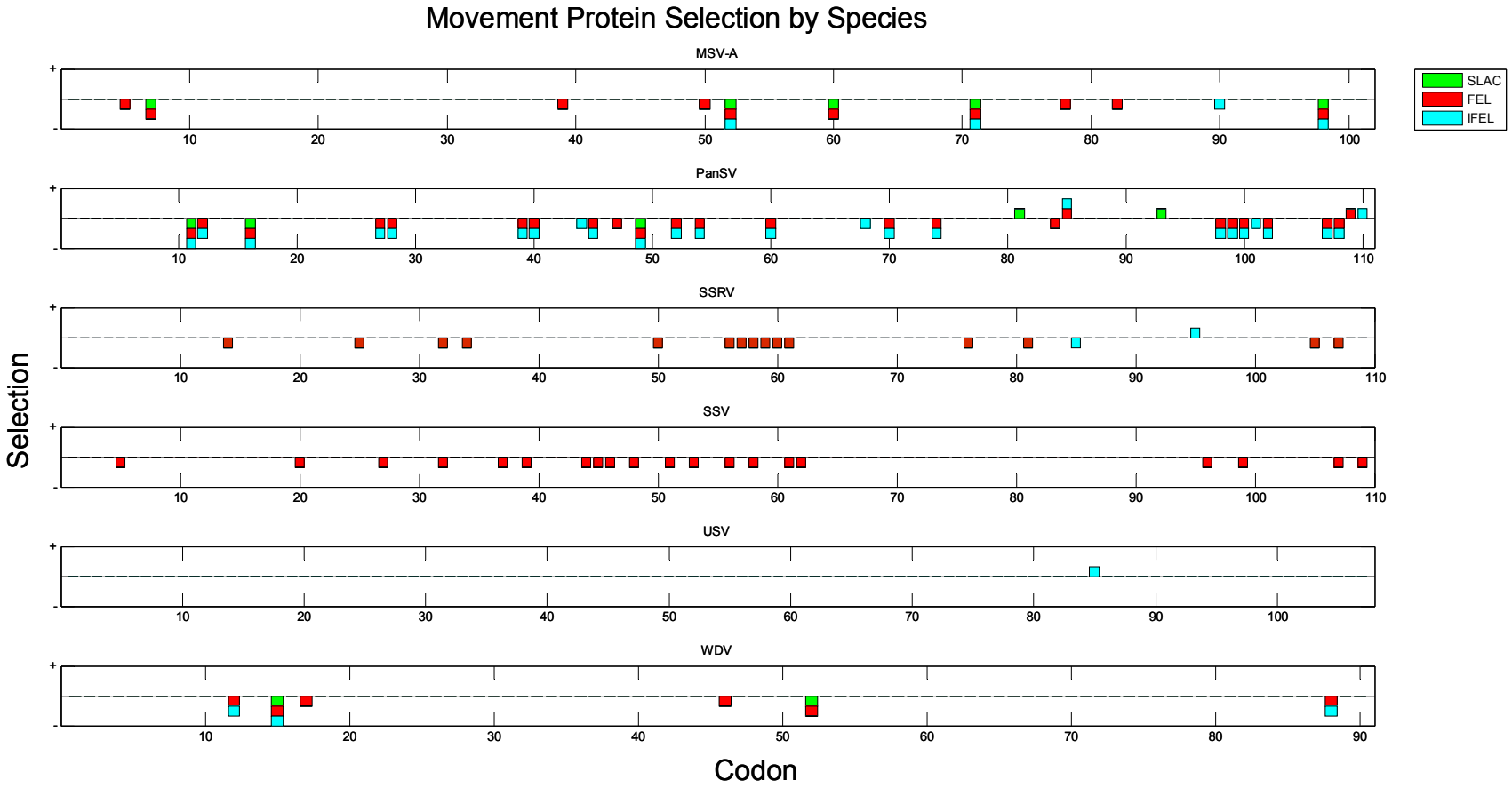


Figure 3.17: Selection analysis on the movement protein by species. Sites under selection as detected by SLAC, FEL and IFEL are presented here. Detection is indicated by coloured boxes, which may be stacked if multiple algorithms detected selection. See figure legend for colour code. An example of movement protein structure is presented above to indicate relevant areas. See the Supplementary material for full alignments.

3.4.6 Selection by Host

To examine the links between evolution and host species selection analysis on the host comparison datasets was carried out.

Coat Protein

Detectable selection was distributed across the coat protein gene and was predominantly negative (Figure 3.18). The panicum dataset exhibited the strongest detectable selection, with SLAC, FEL and IFEL all confirming numerous sites (see Supplementary Material). Conserved regions were again often under negative selection. The **LQIQ** region and **KRKR** residues were partially under negative selection. Sites flanking the **LQIQ** residues were under positive selection in the panicum dataset. Urochloa, setaria and maize had considerably less detectable selection. All datasets contained more detectable selection than 'by species' datasets. The maize and sugarcane datasets notably had detectable positive selection within the nuclear localisation signal, albeit only at a limited number of sites. The C-terminal region was under more selection in this dataset than in the 'by species' datasets.

Movement protein

Selection was detected across the length of the movement protein gene (Fig 3.19). Selection was predominantly negative and tended to be concentrated in the C-terminal and central regions of the protein. Positive selection was predominantly detectable in the C-terminal regions. The setaria and sugarcane datasets both contained one detectable positive site in the N-terminal region. The sugarcane dataset also had the most sites under detectable positive selection. Significant negative selection was detected within the transmembrane domain (Around codons 38-63).

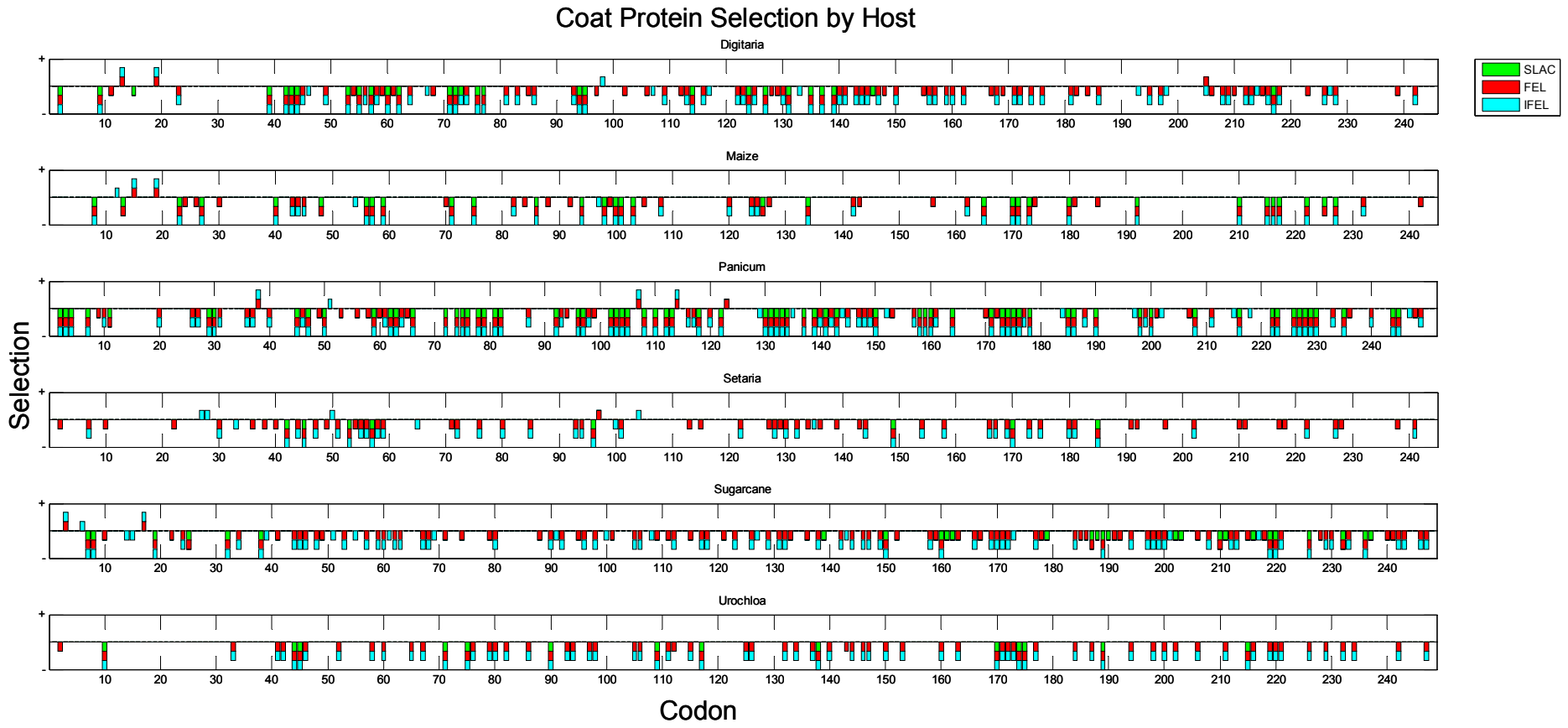


Figure 3.18: Selection analysis on the coat protein by host species. Sites under selection as detected by SLAC, FEL and IFEL are presented here. Detection is indicated by coloured boxes, which may be stacked if multiple algorithms detected selection. Green denotes SLAC, red FEL and blue IFEL. See Supplementary material for full alignments.

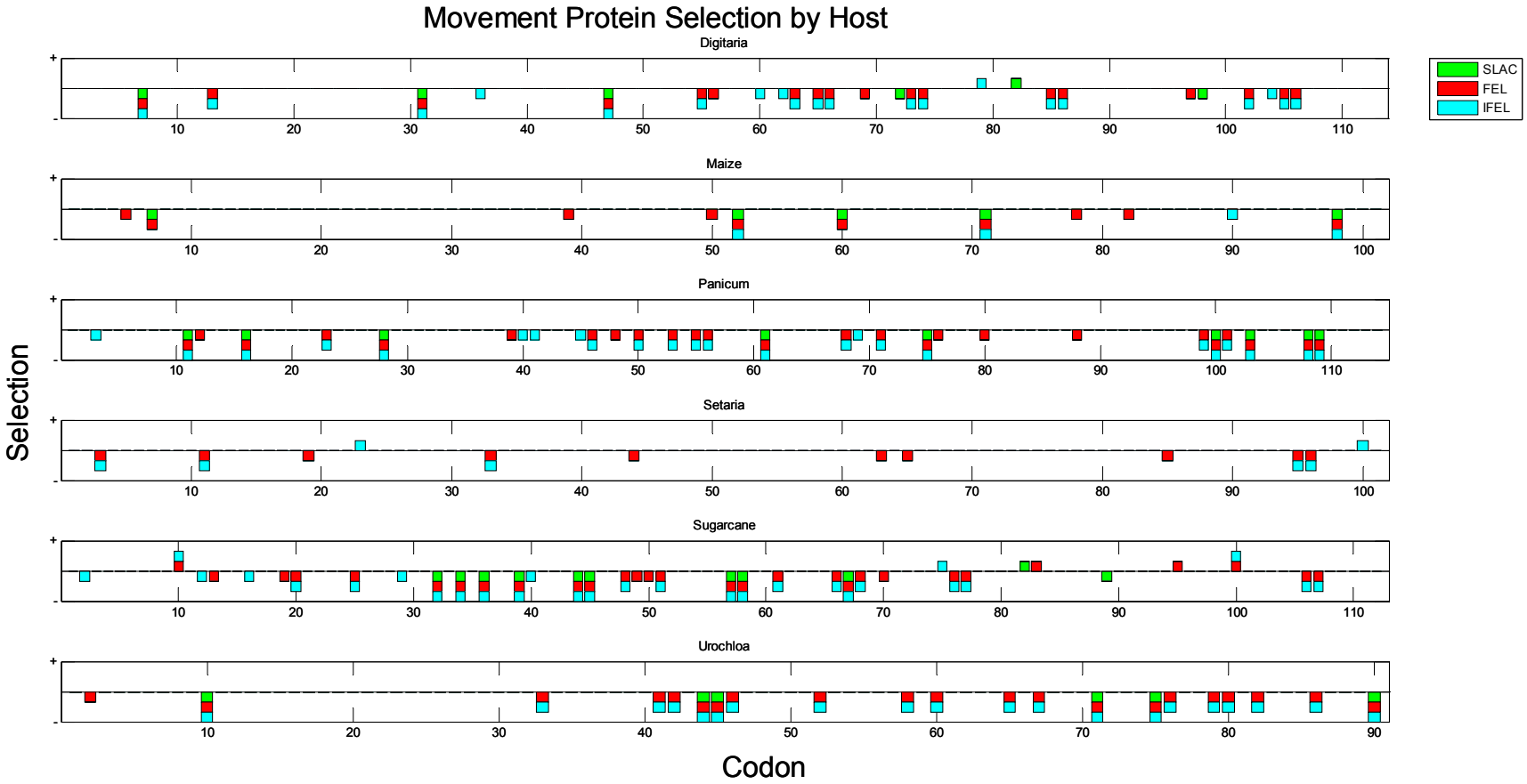


Figure 3.19: Selection analysis on the movement protein by host species. Sites under selection as detected by SLAC, FEL and IFEL are presented here. Detection is indicated by coloured boxes, which may be stacked if multiple algorithms detected selection. Green denotes SLAC, red FEL and blue IFEL. See Supplementary material for full alignments.

The structure of any mastrevirus movement protein has not been solved. Sites under selection in the coat protein gene were modeled to 3D structures of the coat protein (Fig 20, 21). Nucleotide sites under selection were compared to the amino acid translation within their alignment, and any codons under selection were marked. Sites under selection were detectable across the protein, with several clusters being observed around positions 56, 105 and 171 in multiple datasets. A small cluster around codons 124-127 was observed in maize which corresponds to a small alpha helical structure. Negative selection was observable near many of the conserved sites previously identified.

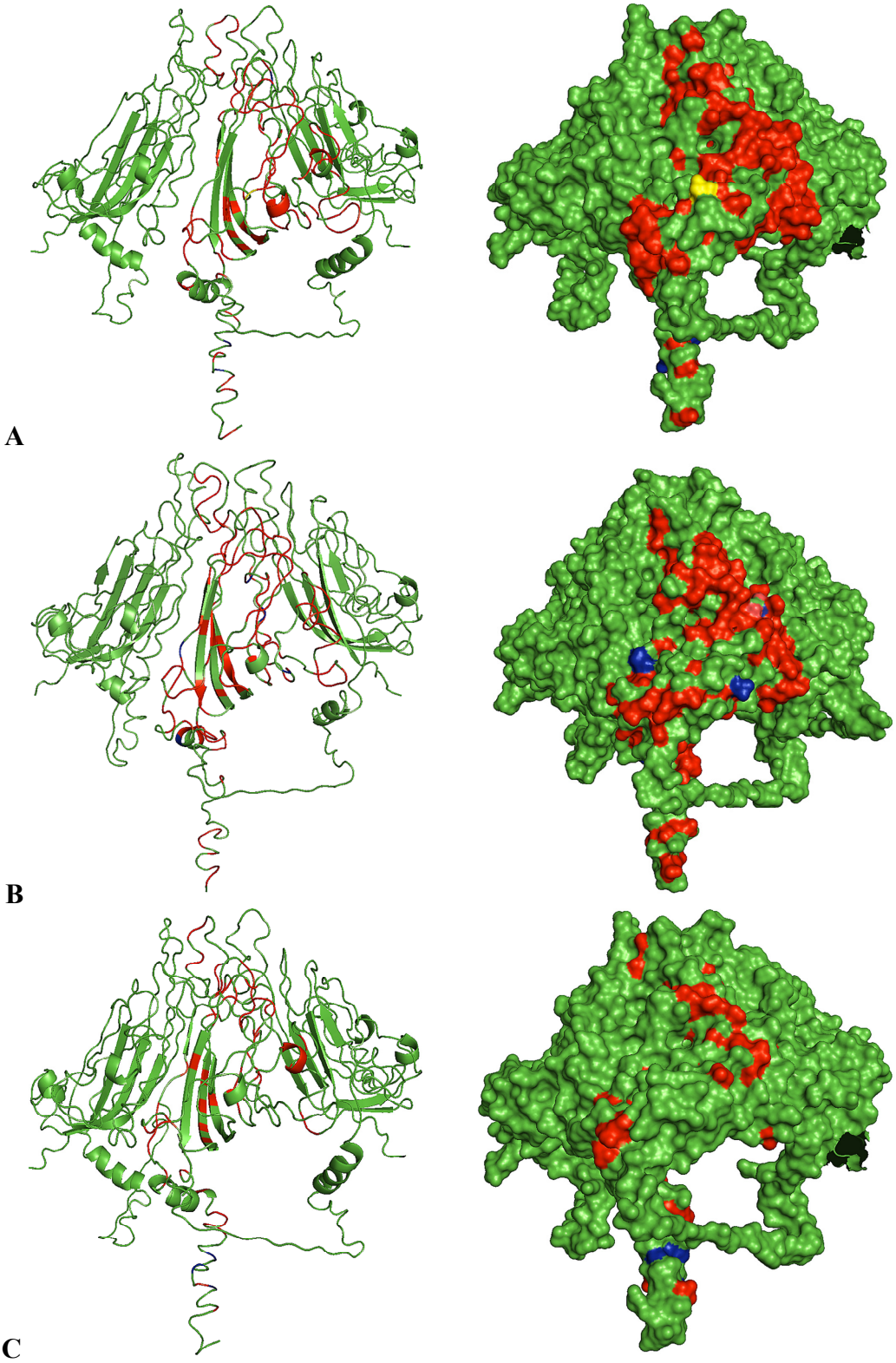


Figure 3.20: 3D models of selection by host. **A**: Digitaria. **B**: Panicum **C.** Maize Red indicates amino acids with codons under negative selection. Blue indicates amino acids with codons under positive selection. Yellow indicates detection of positive and negative selection at that site by different algorithms.

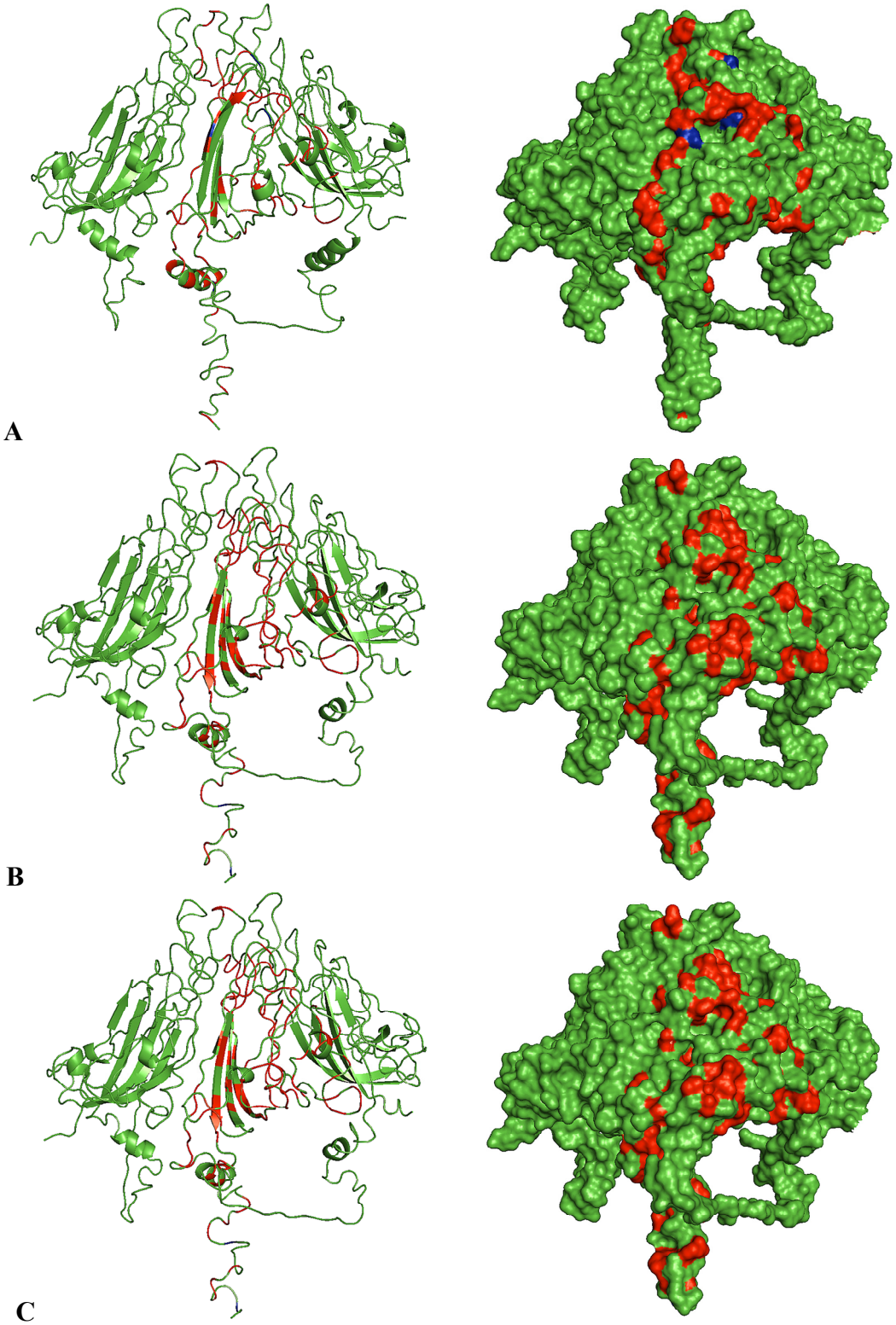


Figure 3.21: 3D Model of selection by host: **A**: Setaria **B**: Sugarcane **C**: Urochloa. Red indicates amino acids with codons under negative selection. Blue indicates amino acids with codons under positive selection. Yellow indicates detection of positive and negative selection at that site by different algorithms.

3.4.7 Codon Usage

In order to determine if codon bias was a significant factor in mastrevirus evolution a codon usage analysis was performed for the coat protein over the aggregated datasets and within the by host subsets. Over the entire dataset a significant preference for certain codons was observed (Table 1). All datasets showed similar trends in codon usage. Notable was the very strong bias towards AAG. GGA for Glycine and CAG for Glutamine were favored highly amongst all datasets. GAC coding for Aspartic acid was highly favored over the entire dataset, as well as in digitaria, maize and setaria, but was not significantly favored in all other datasets. This suggests that there is a preference for certain codons in mastreviruses.

Coat Protein

	Entire	Digitaria	Maize	Panicum	Setaria	Sugarcane	Urochloa		Entire	Digitaria	Maize	Panicum	Setaria	Sugarcane	Urochloa
UUU(F)	3.1(0.75)	3.1(0.76)	3.0(0.75)	2.6(0.56)	2.0(0.50)	3.2(0.67)	2.0(0.47)	**UCU(S)**	2.0(0.68)	1.9(0.64)	1.9(0.65)	2.6(1.05)	1.6(0.54)	4.5(1.62)	4.9(1.56)
UUC(F)	5.2(1.25)	5.1(1.24)	5.0(1.25)	6.6(1.44)	6.0(1.50)	6.4(1.33)	6.6(1.53)	**UCC(S)**	5.0(1.72)	5.6(1.86)	4.8(1.63)	4.8(1.94)	6.4(2.16)	3.8(1.39)	6.6(2.11)
UUA(L)	0.2(0.08)	0.1(0.03)	0.0(0.00)	0.4(0.19)	0.2(0.09)	0.5(0.28)	0.0(0.00)	**UCA(S)**	1.1(0.38)	1.3(0.42)	1.0(0.34)	2.0(0.80)	1.4(0.47)	2.5(0.93)	0.7(0.23)
UUG(L)	2.6(1.20)	2.4(1.10)	2.9(1.34)	4.2(1.87)	2.2(1.03)	2.9(1.51)	2.5(1.13)	**UCG(S)**	2.7(0.93)	2.8(0.93)	3.1(1.06)	1.7(0.66)	2.4(0.81)	1.7(0.63)	2.5(0.80)
CUU(L)	1.3(0.61)	1.5(0.70)	1.1(0.53)	1.7(0.77)	0.6(0.28)	1.3(0.66)	2.1(0.94)	**CCU(P)**	3.5(1.18)	3.7(1.26)	3.7(1.21)	4.9(1.14)	2.8(0.92)	5.8(1.66)	4.9(1.60)
CUC(L)	3.2(1.47)	2.7(1.26)	3.6(1.65)	3.0(1.36)	3.4(1.59)	2.5(1.28)	4.4(1.97)	**CCC(P)**	2.9(0.97)	2.4(0.81)	3.6(1.15)	4.5(1.05)	2.4(0.79)	2.5(0.73)	2.7(0.89)
CUA(L)	0.8(0.35)	0.7(0.31)	0.7(0.33)	0.2(0.10)	1.8(0.84)	0.6(0.33)	0.4(0.16)	**CCA(P)**	2.1(0.72)	2.0(0.66)	2.0(0.65)	1.8(0.43)	1.8(0.59)	1.6(0.47)	1.1(0.38)
CUG(L)	4.9(2.30)	5.7(2.60)	4.7(2.15)	3.8(1.70)	4.6(2.16)	3.7(1.94)	4.0(1.81)	**CCG(P)**	3.4(1.14)	3.7(1.26)	3.1(0.99)	6.0(1.39)	5.2(1.70)	4.0(1.14)	3.4(1.13)
AUU(I)	3.2(0.91)	3.2(0.95)	2.8(0.79)	4.6(1.26)	2.4(0.78)	5.4(1.58)	4.4(1.26)	**ACU(T)**	5.1(0.84)	5.0(0.80)	5.2(0.84)	5.2(1.08)	6.6(1.06)	5.8(1.10)	5.4(1.01)
AUC(I)	4.9(1.40)	3.8(1.12)	5.9(1.65)	4.9(1.35)	4.2(1.37)	3.7(1.10)	3.4(0.97)	**ACC(T)**	9.4(1.55)	9.6(1.56)	10.1(1.64)	6.3(1.31)	8.2(1.32)	6.8(1.29)	7.4(1.41)
AUA(I)	2.4(0.68)	3.1(0.92)	2.0(0.56)	1.4(0.39)	2.6(0.85)	1.1(0.32)	2.6(0.77)	**ACA(T)**	3.2(0.52)	3.7(0.59)	2.9(0.47)	1.8(0.38)	4.6(0.74)	2.9(0.55)	2.2(0.42)
AUG(M)	5.2(1.00)	5.2(1.00)	5.0(1.00)	3.5(1.00)	4.8(1.00)	3.6(1.00)	4.4(1.00)	**ACG(T)**	6.6(1.10)	6.4(1.04)	6.5(1.05)	5.8(1.23)	5.4(0.87)	5.6(1.06)	6.1(1.16)
GUU(V)	3.4(0.84)	2.5(0.67)	3.1(0.88)	7.2(1.42)	3.4(0.87)	5.4(1.01)	2.8(0.63)	**GCU(A)**	4.6(1.00)	4.8(1.04)	4.2(0.87)	6.8(1.32)	4.6(1.02)	5.7(1.07)	6.6(1.24)
GUC(V)	3.3(0.83)	3.8(1.03)	2.3(0.64)	5.4(1.07)	4.0(1.03)	4.5(0.84)	5.9(1.33)	**GCC(A)**	8.2(1.76)	7.3(1.56)	10.0(2.09)	10.2(1.96)	6.6(1.47)	10.5(1.96)	5.8(1.08)
GUA(V)	3.0(0.75)	2.5(0.69)	3.0(0.84)	0.4(0.08)	1.8(0.46)	1.8(0.34)	3.2(0.73)	**GCA(A)**	2.7(0.57)	3.1(0.67)	2.0(0.41)	0.4(0.07)	3.6(0.80)	2.3(0.43)	4.6(0.85)
GUG(V)	6.4(1.58)	5.9(1.61)	5.8(1.64)	7.2(1.44)	6.4(1.64)	9.6(1.81)	5.8(1.31)	**GCG(A)**	3.1(0.67)	3.4(0.72)	3.0(0.63)	3.3(0.64)	3.2(0.71)	2.9(0.54)	4.4(0.83)
UAU(Y)	2.9(0.63)	2.6(0.60)	1.9(0.42)	1.5(0.34)	2.4(0.57)	3.5(0.76)	2.2(0.47)	**UGU(C)**	2.2(1.04)	2.0(0.97)	2.0(1.01)	1.3(0.36)	2.8(1.33)	2.2(0.81)	1.9(0.90)
UAC(Y)	6.3(1.37)	6.2(1.40)	7.1(1.58)	7.4(1.66)	6.0(1.43)	5.6(1.24)	7.2(1.53)	**UGC(C)**	2.0(0.96)	2.2(1.03)	2.0(0.99)	5.8(1.64)	1.4(0.67)	3.2(1.19)	2.3(1.10)
UAA(*)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	**UGA(*)**	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)
UAG(*)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	**UGG(W)**	7.3(1.00)	6.9(1.00)	7.0(1.00)	6.0(1.00)	7.0(1.00)	5.5(1.00)	6.4(1.00)
CAU(H)	2.7(0.95)	2.6(0.87)	3.0(1.00)	1.7(0.56)	2.0(0.67)	3.2(1.08)	2.3(0.76)	**CGU(R)**	1.7(0.61)	1.4(0.56)	1.2(0.41)	2.8(1.00)	2.0(0.79)	5.5(1.82)	1.4(0.51)
CAC(H)	3.0(1.05)	3.4(1.13)	3.0(1.00)	4.3(1.44)	4.0(1.33)	2.7(0.92)	3.7(1.24)	**CGC(R)**	3.5(1.28)	3.2(1.24)	4.0(1.40)	4.3(1.56)	2.4(0.95)	1.7(0.57)	3.1(1.10)
CAA(Q)	2.7(0.61)	2.7(0.60)	3.0(0.66)	1.0(0.33)	1.2(0.28)	0.5(0.15)	1.9(0.56)	**CGA(R)**	2.1(0.77)	2.5(0.99)	1.9(0.66)	1.2(0.41)	2.2(0.87)	1.5(0.48)	1.0(0.36)
CAG(Q)	6.2(1.39)	6.4(1.40)	6.0(1.34)	5.0(1.67)	7.4(1.72)	5.6(1.85)	4.8(1.44)	**CGG(R)**	4.7(1.72)	3.9(1.54)	5.1(1.80)	3.6(1.28)	4.6(1.82)	5.0(1.64)	3.9(1.38)
AAU(N)	4.6(0.85)	3.8(0.73)	4.1(0.81)	2.3(0.65)	3.8(0.76)	5.4(1.20)	2.8(0.70)	**AGU(S)**	3.2(1.11)	3.1(1.04)	3.2(1.07)	2.5(0.99)	3.6(1.21)	1.9(0.70)	2.4(0.78)
AAC(N)	6.3(1.15)	6.7(1.27)	6.0(1.19)	4.8(1.35)	6.2(1.24)	3.5(0.80)	5.1(1.30)	**AGC(S)**	3.5(1.19)	3.3(1.10)	3.7(1.25)	1.4(0.56)	2.4(0.81)	2.0(0.73)	1.6(0.53)
AAA(K)	3.0(0.33)	2.9(0.32)	2.3(0.27)	1.3(0.14)	3.8(0.40)	2.0(0.25)	2.4(0.26)	**AGA(R)**	1.7(0.62)	1.9(0.76)	2.0(0.71)	0.7(0.26)	1.4(0.55)	1.2(0.39)	2.5(0.90)
AAG(K)	15.4(1.67)	15.6(1.68)	14.7(1.73)	16.8(1.86)	15.4(1.60)	14.2(1.75)	15.5(1.74)	**AGG(R)**	2.7(1.00)	2.3(0.90)	2.9(1.02)	4.2(1.49)	2.6(1.03)	3.4(1.10)	4.9(1.74)
GAU(D)	3.6(0.59)	3.0(0.52)	3.4(0.57)	4.0(0.82)	3.4(0.54)	5.0(0.96)	4.1(0.76)	**GGU(G)**	5.7(0.97)	5.5(0.91)	4.0(0.68)	4.6(0.74)	7.8(1.27)	7.9(1.27)	6.3(0.96)
GAC(D)	8.7(1.41)	8.6(1.48)	8.6(1.43)	5.7(1.18)	9.2(1.46)	5.4(1.04)	6.6(1.24)	**GGC(G)**	4.6(0.78)	5.2(0.87)	4.5(0.76)	11.3(1.82)	4.6(0.75)	7.7(1.24)	8.4(1.27)
GAA(E)	0.6(0.27)	0.4(0.20)	0.0(0.00)	1.1(0.28)	0.6(0.24)	2.0(0.57)	0.6(0.18)	**GGA(G)**	9.0(1.52)	9.3(1.54)	10.7(1.79)	4.3(0.69)	6.0(0.98)	4.5(0.71)	6.4(0.97)
GAG(E)	3.8(1.73)	4.0(1.80)	4.0(2.00)	6.6(1.72)	4.4(1.76)	5.0(1.43)	5.9(1.82)	**GGG(G)**	4.3(0.73)	4.2(0.69)	4.6(0.78)	4.6(0.74)	6.2(1.01)	4.9(0.79)	5.3(0.80)

Table 3.1: Codon distribution amongst various coat protein datasets. The entire dataset is used for the first column. All other columns are from the datasets arranged by host species. Red highlights indicate codons that are favoured highly in this dataset.

3.5 Discussion

Mastreviruses are a significant contributor to crop damage worldwide. However a detailed study of mastrevirus evolution across all species has not been carried out. Here an analysis of approximately 400 mastrevirus gene sequences was undertaken, identifying movement and coat protein regions by sequence similarity, and scanning these regions for selection. The aim of this study was to use this data to infer sites within these functional regions that are of evolutionary importance and then to attempt to infer internal or external influences on evolution by comparison of host grouped and non host grouped datasets.

3.5.1 Sequence alignment and Conservation

In this study numerous sequences were aligned using CLUSTAL W (Thompson *et al*., 1994). Sequence alignment is a critical part of bioinformatics studies, allowing both detecting of key sequences by homology comparison and constructing a dataset in which relevant sequences can be accurately compared. CLUSTAL W was chosen as it provides a reliable sequence alignment across even widely divergent sequences and in datasets with a large number of sequences. Alignment accuracy is essential to proper conservation and selection analysis, as misaligned sequences will compare dissimilar information and lead to false conclusions. Owing to the main focus of this study being selection at the protein level, all sequences were aligned at such, to further enhance alignment accuracy and increase the chance of comparing relevant regions.

By comparison of homologous regions MP and CP were identified across a range of mastreviruses. Functional sites similar to those previously detected in MSV (NLS, DNA binding region and hydrophobic domains) were identified. Highly conserved regions in these sites were also found. The **KRKR** region flanking the NLS was conserved in a large number of mastreviruses. As observed in earlier studies basic sites around this region were widely conserved, again suggesting that the basic nature of the CP aids its DNA binding ability (Liu *et al.,* 1997a, 1999). The **LQIQ** domain was also widely conserved. This domain is positioned at the base of the N-terminal arm structure, and

could therefore be responsible for positioning of arm. Few regions were absolutely conserved across all mastreviruses, however the C-terminal end of the protein was almost absolutely conserved with the exception of WDV, MiSV, BeYDV and DSV. This combined with the high conservation along the lower parts of the protein in many datasets may suggest some important function of this region, such as interaction with other capsomers.

Within the movement protein sites were widely conserved in both the species and strain datasets, but less conserved over all mastreviruses. Numerous conserved sites were detected in the by host datasets, particularly in the transmembrane domain, suggesting that this region is important to function in particular hosts.

The MSV- A-K and MSV -Kom/-Set analyses provide possible target regions for studies on virulence. MSV-A is the strain that is highly virulent in maize, whilst B and C are rarely found in maize but common in other grasses (e.g. digitaria). Variable sites in the sequence may potentially account for these differences in virulence phenotypes. While it is impossible to say from the data which regions are specifically involved, the high concentration of variable sites within the DNA binding region of the MSV-Kom/-Set may warrant future investigation. The two variable clusters in the movement protein are interesting, as they are each positioned on the opposite side of the potential trans-membrane domain and could potentially have roles in efficient inter-cell transport. Alternatively they may simply enhance the innate toxicity of the movement protein observed in other studies (Hou *et al.,* 2001), leading to the increased virulence seen in MSV-Kom.

3.5.2 Selection

Contemporary selection analysis relies upon estimation of dN/dS rates, or the rate of synonymous substitution versus the rate of non synonymous substitution at an amino acid site. dN/dS ratios greater than one indicate positive selection, less than one indicates purifying selection and one is selectively neutral. Various methods are available for selection analysis, each with its advantages and disadvantages. In this study a number of methods were used in order to gain an accurate representation of actual selection. Due to

the required amount of computational time required for selection analysis the Datamonkey implementation of the HyPhy toolkit was used for this study (Pond *et al.,* 2005a, Pond & Frost, 2005b,). It offers model testing capability and SLAC, FEL, IFEL and REL for selection detection.

Model testing for evolutionary studies is critical to ensuring accurate estimation of substitution rates (Posada & Buckley, 2004). Although no model can be entirely accurate, simulating the most likely nucleotide substitution model is important to ensure consistent results. The HyPhy package implements a three stage model selection process utilising heirarchical selection, followed by nested model selection, followed by comparison to the Akaike information criterion. This ensures selection of the most accurate possible method of nucleotide substitution. SLAC, REL, IFEL and REL tests can then be run. REL was not used in this study as it can potentially generate false positives in small datasets (Pond & Frost, 2005c).

SLAC, the most conservative of these methods operates by estimating ancestral sequences based on phylogeny and the chosen substitution model and then estimating the number of changes that have occurred (Pond & Frost, 2005c). SLAC gives the least false positives, at the expense of missing some true positives. FEL and IFEL are similar, and are less conservative, at the risk of increased false positives. IFEL differs only in that it only measures substations along internal branches, allowing discrimination of population level effects (Pond *et al*., 2006).

Using these methods selection analysis was performed across all available mastrevirus groups. Selection within the coat protein showed significant purifying selection along the length of the protein. Purifying selection was detected in multiple sites across the DNA-binding region in all datasets except USV. Purifying selection was expected in functional sites / domains, as the majority of changes to these sites / domains are likely to have deleterious results. Selection detectable in other regions is unsurprising, particularly in light of the high conservation of the C-terminal region across highly divergent mastreviruses. However purifying selection was not detected in several absolutely

conserved regions. It has been suggested that population level effects amongst a species may interfere with selection analysis, particularly in microorganisms (Kryazhimsky & Plotkin., 2008). Were this the case it would be expected that the selection by host dataset, in which numerous relatively divergent species are present, would show evidence of selection across these regions. The IFEL test should also be capable of detecting selection within this type of dataset (Pond *et al*., 2006). The most likely explanation for the low level of selection detectable at these sites is that these regions are absolutely necessary for virus function. Viruses without these regions are likely to not be sampled or be underrepresented, making it possible that selection can not be detected at these sites. It is also possible that the number of isolates present within certain datasets provided insufficient resolution for detection of all selected sites. Regardless, the high conservation and low level of negative section present in this region is notable, as a large amount of the CP region maps to a recombination cold spot (Varsani *et al*., 2008a). Recombination is a predominant process in mastrevirus evolution, and the absence of recombination and limited levels of purifying selection suggest neutral evolution to be the dominant factor in the evolution of this region. This supports a recent study suggesting that mastreviruses are predominantly evolving under genetic drift as well as providing evidence against the codivergence hypothesis which requires purifying selection at over 99% of sites (Harkins *et al*., 2009a, Wu *et al*., 2008).

Sites under selection did vary within each dataset, although several areas that were predominantly selected were detected. This may indicate regions that are significant to particular virus strains or species. The positive selection detectable on the NLS of the MSV-A dataset is particularly significant as this is a crucial functional region. Charged sites (and flanking regions) under selection varied between datasets, suggesting these minor variations are significant to each species characteristics. Charged regions are particularly important within the CP as positively charged sites contribute to DNA binding function.

The movement protein had considerably less detectable selection than the coat protein. This was partially expected as it has the least known inter-protein/inter-gene interactions

of all mastrevirus coding regions (Martin *et al*.. 2005). Maintenance of regions responsible for these interactions is a significant factor in mastrevirus evolution, and the lack of them may be less restrictive on variation within this region. The lack of known functional sites makes it difficult to comment on the relevance of the selection detected in this study, however the high levels of selection detectable in C-terminal regions may suggest an important structural or functional activity.

3.5.3 Selection by Host

The selection by host analysis showed much stronger selection signals, than the 'by species' datasets.. Positive selection in these datasets was also more pronounced, and was frequently located within the DNA binding domain or NLS. Central regions flanking the DNA domain showed positive selection in the panicum dataset. Digitaria and maize had positively selected sites within the NLS, possibly suggesting that they are undergoing adaptation to their host in this region.

The movement protein shows increased detectable selection in the by host datasets. Selection within the transmembrane domain was predominantly negative as expected for a functional region. Numerous negatively detected sites were present between the C-terminal region of the MP and the transmembrane domain, suggesting this region may be significant to activity. Several sites in this region were under positive selection in the sugarcane datasets, potentially indicating host adaptation occurring within this region.

3.5.4 Codon Usage

Codon usage data indicates a strong preference for certain codons in the coat protein. These may reflect favoured tRNA types within plant species, and the data reflects this to some extent, with datasets arranged by species showing deviation from the codon usage seen in the entire dataset. This could indicate host codon bias is a factor in mastrevirus evolution / host adaptation. Codon bias can provide constraints upon variability in neutral sites, and may help explain the existence of conserved unselected regions in mastrevirus genomes (Ermolaeva, 2001). Codon bias is thought to take effect by limiting the speed of transcription for rare codons, thus reducing the availability of required proteins where rare codons are used. It is thus favourable for a virus, which uses the host's replicative

machinery, to match the host species codon usage. Codon usage can also be attributable to GC content of the genome, and as such may not have specific significance.

3.5.5 Implications of Selection and Conservation Patterns

A considerable amount of data on the selection and conservation patterns prevalent within mastrevirus genomes was gathered. Although negative selection was detected across both proteins, the majority of sites were selectively neutral. The lack of detailed information available about specific mechanisms of interaction of the coat and movement proteins hinders our ability to conclusively link the selection and conservation data to key motifs within these regions. However, the data reinforces earlier findings on the importance of the N-terminal region of the CP and the NLS. The strong negative selection detectable at many sites along the length of this region suggests that numerous small interactions across this region are more likely to be responsible for subtle variations in CP function than any one obvious key motif. Positive selection in the NLS may indicate it is particularly vital to adaptation to certain host species. Small changes in amino acids and in regions flanking basic sites seem likely to contribute to these variations. Several conserved sites, such as the C-terminal region were not under significant negative selection, suggesting that they may be absolutely required in mastreviruses. Small variations in these sites may thus be determinants of some key activity or structure within the protein. Sites under selection but not absolutely conserved within these non-functional regions are likely to be relevant to protein structure, and as such may vary between species due to adaptation to different host plants.

The implications of movement protein selection and conservation are more difficult to identify. Selection within the C-terminal region seems to suggest an important function for this region that may be relevant to infectivity within a particular host. The negative selection in the transmembrane domain may indicate sites responsible for minor variation in function of this region and may bear further investigation. Highly conserved sites in this region in the by host datasets may indicate specific conformations are important in allowing this regions function in particular hosts.

4 Selection and conservation patterns in the mastrevirus replication protein.

4.1 Abstract

Mastrevirus replication protein is responsible for the initiation of replication in mastreviruses. Numerous binding sites and motifs have been identified on the Rep protein by earlier homology studies. The wealth of mastrevirus sequences now available make large scale homology studies possible. The aim of this research was to identify homologous sequences and sequences under selection within these sequences, to attempt to identify vital regions in the Rep protein sequence. Using sequence alignment techniques and tools implemented in HyPhy software several key conserved regions in the Rep protein were identified. Codons under selection were identified throughout the replication protein. Several functional sites were under detectable selection and two important regions between the RCR motifs were identified. Despite this the protein was overall selectively neutral. An interaction between several motifs within the catalytic domain was identified, and may have implications for the function of this region.

4.2 Introduction

Geminiviruses are a diverse group of ssDNA viruses that infect a range of plants. The genus *Mastrevirus* comprises the second biggest group of Geminiviruses. All mastrevirues replicate via rolling circle replication. This process requires only the Replication protein (Rep) to function (Gutierrez., 2000). Rep contains a number of binding sites controlling this function. Three of these, rolling circle replication motifs one to three (RCR I-III), are responsible for the nicking process that is responsible for initiation of mastrevirus replication (Fig 4.1). The 3D structure of the catalytic domain containing these RCR motifs has been solved for *tomato yellow leaf curl virus* (TYLCV), a member of the genus *Begomovirus* (Campos-Olivas *et al*., 2002).

Rep contains a number of other motifs that have been shown to have various functions. Rep contains a retinoblastoma binding motif (Rbr) in the C-terminal region which is likely to interact with the cell cycle of plant hosts, however, it is only functional in RepA (Xie *et al*., 1995). The 37 C-terminal amino acids of Rep have been shown to bind to GRAB proteins, which suppress mastrevirus activity (Xie *et al*., 1995). dNTP binding regions are also found in the C-terminal domain (Gorbalenya and Koonin, 1989).

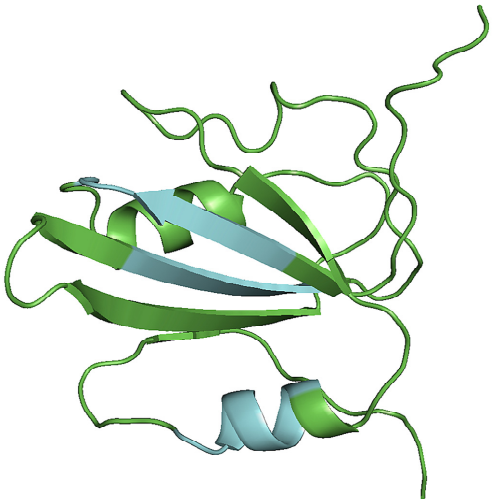


Figure 4.1: Catalytic domain of Rep protein, showing RCR motifs in cyan. RCR I and II are on the beta sheet, RCR III on the alpha helix.

Mastrevirus evolution has become has come into the limelight in recent years. Estimates of the basal rate of mutation in mastreviruses have placed it as high as $2x10^{-3}$ mutations per site per year (van der Walt *et al*., 2008). The exact nature of evolution has been controversial, with two key theories of mastrevirus evolution being suggested. One model suggests that codivergance is the primary model of mastrevirus evolution (Wu *et al*., 2008). The other suggests that neutral evolution is the primary model of mastrevirus evolution (Harkins *et al*., 2009a). The codivergance hypothesis relies on strong negative selection across the mastrevirus genome.

To identify negative selection accurate estimation of relatedness within target regions is essential. Some modern selection detection algorithms rely on drawing a phylogenetic tree and reconstructing ancestral character (Pond *et al*., 2005a). As relatedness within different sections of mastrevirus genomes can be different to overall sequence identity due to processes such as recombination, accounting for this is important. A maximum likelihood (ML) phylogenetic relationship of all mastreviruses of the replication protein is provided in Figure 4.2.

This study utilises a series of experiments involving a wide range of mastreviruses. Using homology modeling, sequence conservation and selection algorithms a number of conserved sequences and regions under selection were identified. These results show insufficient negative selection is detectable on the mastrevirus replication gene to support the codivergance hypothesis given the high basal mutation rate of mastreviruses. Similar to the results found by Harkins *et al,* (2009a) the results suggest mastreviruses evolve under a predominantly neutral selective model. A number of conserved regions were also identified.

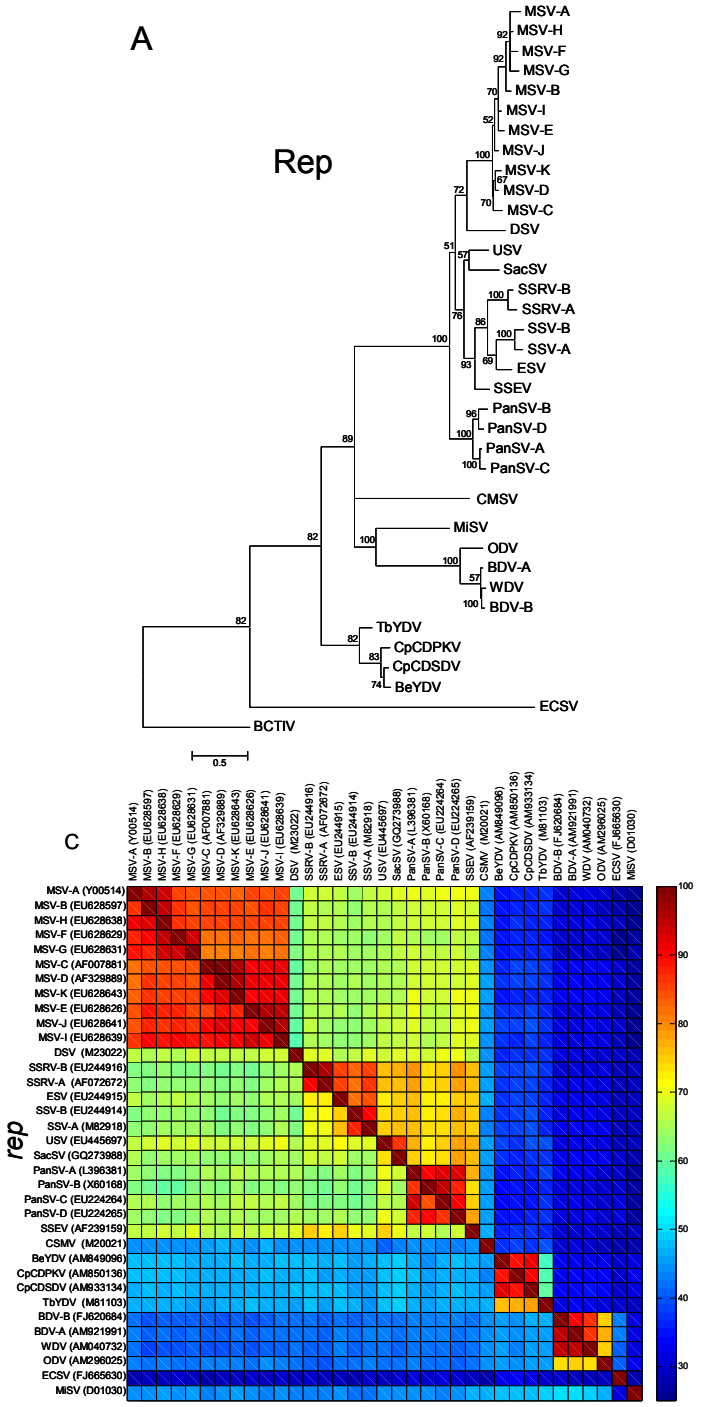


Figure 4.2: Maximum likelihood (ML) phylogenetic relationships based upon alignments of the predicted amino acid sequences of Rep (**A**) and CP (**B**). The ML trees were constructed using PHYML (Guidion and Gascuel, 2003) (best fit Rep and CP models = LG as determined by PROTEST (Abascal *et al*., 2005). Numbers associated with tree branches are indicative of the percentage of 1000 full maximum likelihood bootstrap Replicates supporting the existence of the branches. (C) Two-dimensional graphical Representation of pairwise amino acid sequence identities (calculated with pairwise deletion of gaps; scale Represents percentage identity) of the predicted Rep and CP of Representative mastrevirus species and strain.

4.3 Methods

Sequence Analysis

DSV, MiSV, BeYD and WDV sequences were obtained from the GenBank database. Other unpublished sequences (MSV-A-K, PanSV, SSRV, SSV) were obtained from the MSV (Martin, Shepherd and Varsani) research group.

Sequences were partitioned 'by species' and then aligned using Clustal W (Thompson *et al*., 2004), followed by manual alignment in Mega 4.0 (Tamura *et al*., 2007). Open reading frames were identified by homology to known sequences on the BLAST database. Full length replication protein (the resulting product of splicing between the Rep and RepA transcript) was identified and conserved sequences were recorded. Sequences were then reorganised into a second dataset, arranged by host species, if 3 or more sequences were available.

Codon Selection Analysis

Datasets were realigned at the protein level, and reconverted to a nucleotide FASTA file. Stop codons were then removed to conform with HyPhy specifications. Datasets were then uploaded to the Datamonkey web server which implements the HyPhy bioinformatics package (Pond *et al.,* 2005a, Pond & Frost., 2005b). Each dataset was then run through a model testing procedure to select the most appropriate evolutionary model. GARD analysis was then run in order to account for the effects of recombination. Single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL) and internal fixed effects likelihood methods were then used to detect positive and negative selection (Pond & Frost, 2005c, Pond *et al*., 2006).

3D Modeling

A 3D model of the catalytic domain of Rep from tomato yellow leaf curl virus has been solved (Campos-Olivas *et al*., 2002). This model was used as a template for 3D modeling using the MODELLER (Eswar *et al*., 2006) toolkit implemented through the Easy Modeller GUI. PYMOL (Delano, unpublished) was then used to display space fill models

and cartoon forms of these models. Conserved sites and sites under selection were then mapped onto these in an attempt to illustrate links between likely functional and non-functional regions under selection.

4.4 Results

Sequence Conservation of Mastrevirus Replication Protein

Conserved sequences can be an indicator of functional and structural regions particularly when a 3D model is available. This allows identification of key regions within the 3D structure, as well as identification of the nature of the 3D structure itself. To identify these regions conserved sites within the Rep protein were analysed, and conserved regions within the catalytic domain were mapped to 3D structures. Within the catalytic domain variable sites were slightly more prevalent in the random loop structures and alpha helices than in the beta sheets. Only 43 residues were conserved across the entire mastrevirus dataset, and 8 of these mapped to the catalytic domain (Fig 4.3).

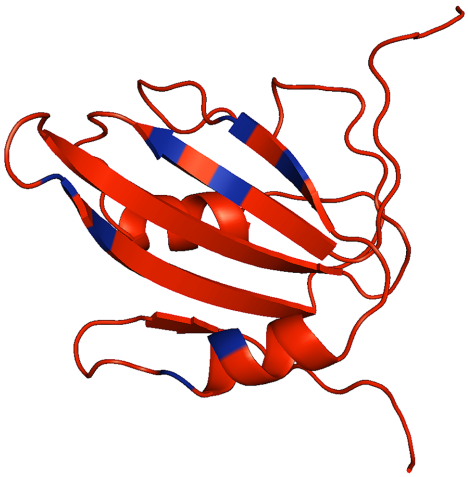


Figure 4.3: Conserved sequences within the catalytic domain of the Rep protein for all mastreviruses. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1.

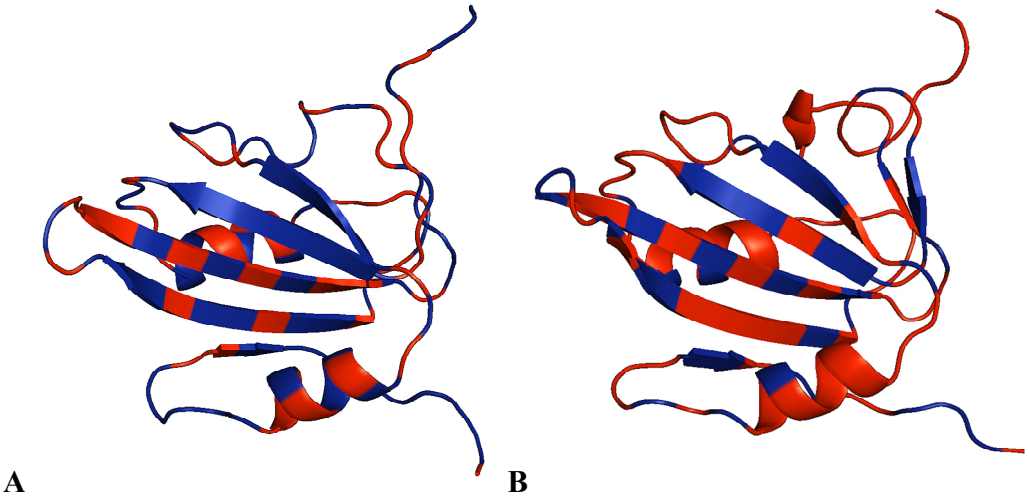


Figure 4.4: **A**. MSV-A-K conservation. **B**: SISV conservation. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

Conserved sites between different groups were then examined. The dataset was split into MSV-A-K strains and the Sugarcane infecting streak virus group, based on the Rep phylogenetic tree (Fig 4.2). 222 of the 346 sites were conserved in the MSV-A-K dataset. The catalytic domain and rolling circle Replication (RCR) domain was highly conserved. The least conserved regions were at the C-terminal end of the catalytic domain, N-terminal of the oligomerisation domain and at the C-terminal of the protein.

The sugarcane infecting streak viruses (SISV) were highly conserved in the catalytic and RCR binding domains. The dNTP binding domains were also conserved in this group. A 3 amino acid region, **NIQ**, was conserved across all SISVs and is situated approximately 7 residues towards the N-terminus from the third RCR motif. The retinoblastoma binding motif (Rbr) was not highly conserved in this group.

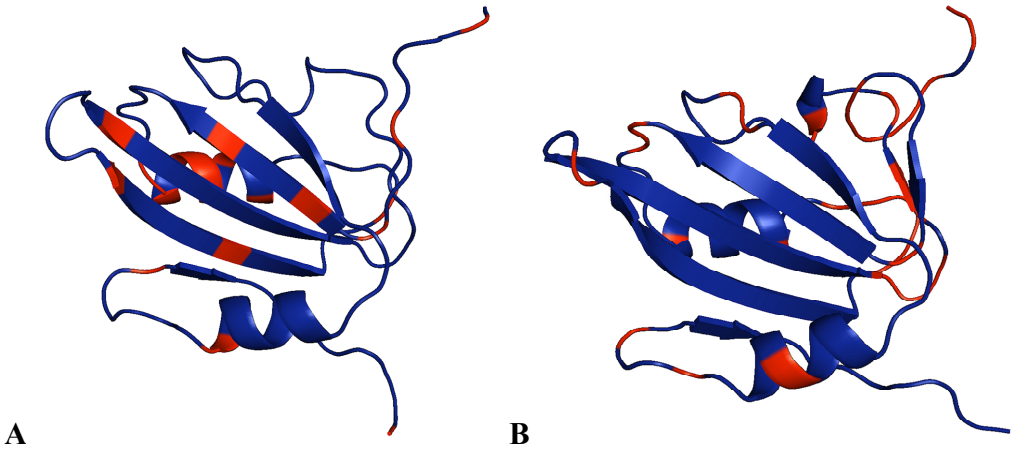


Figure 4.5: A. MSV-complete dataset conservation **B.** PanSV conservation. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

Datasets were then subdivided into groups of host species or viral strain, depending on the number of available sequences. In the MSV-A data 295 of the 361 sites were conserved in the alignment (See Supplementary material). Conserved residues were distributed evenly throughout this dataset. The most highly conserved regions were within the RCR domain, with the entire stretch from RCR motif 2 to RCR motif 3 being conserved. The Rbr motif was not conserved in this dataset, instead it showed a **LL**xx**E**x pattern. The remainder of the dataset showed only minor variations which are difficult to analyse.

The PanSV dataset was conserved over 251 of the 364 sites. The N-terminal region of the catalytic domain differed from that of the MSV set, having the sequence **EGRH** as the predominant sequence rather than **SNRQ**. The catalytic domain was again highly conserved, with a considerably conserved region between the RCR motifs one and two. The C-terminal region of the protein was considerably less conserved than that of the MSV dataset. The oligomerisation domain was also highly conserved within the PanSVs.

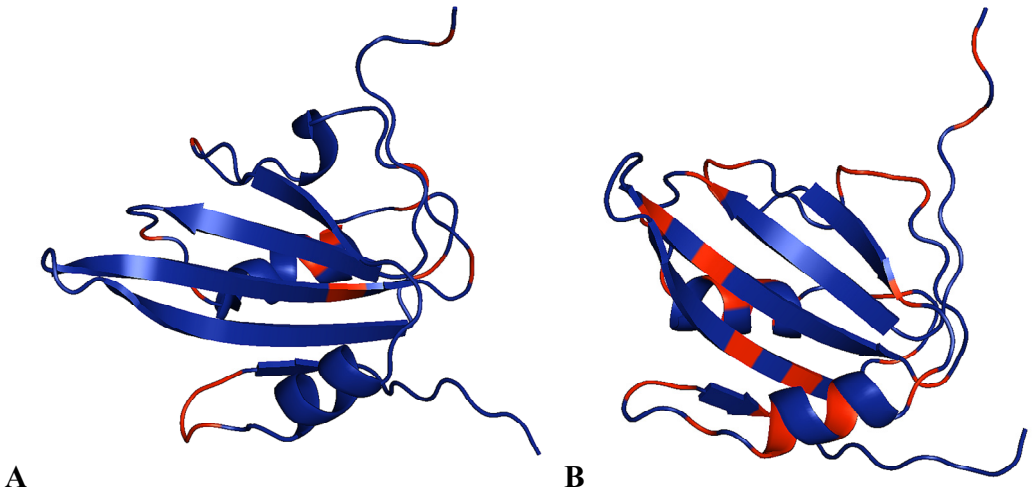
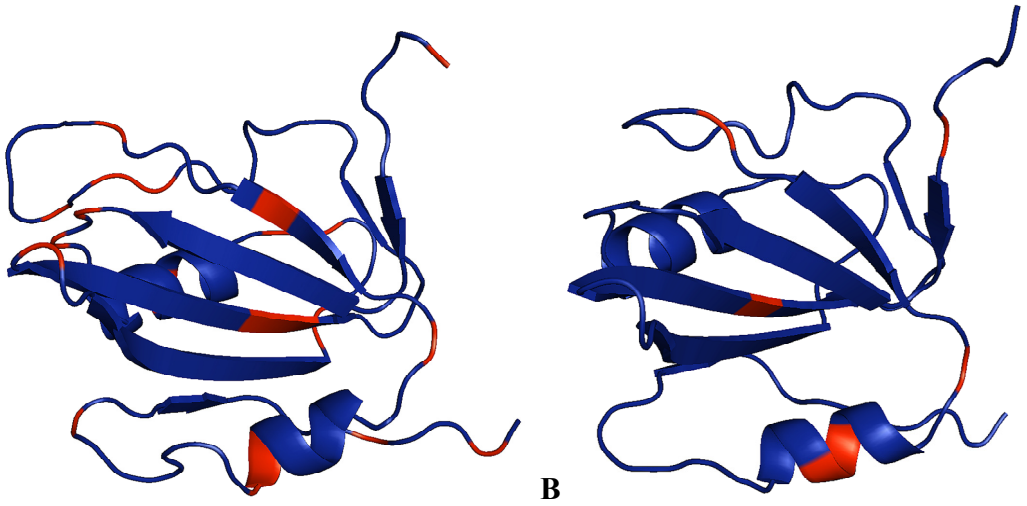


Figure 4.6: A. SSRV conservation. **B.** SSV conservation. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

The SSRV set was conserved over 320 of the 362 sites (Fig 4.6). This high conservation is likely due to the small number of SSRV sequences available. C-terminal regions were less conserved than the bulk of the protein. A region situated between RCR motifs one and two were also less conserved. The SSV set was again limited by small sequences number. 300 of the 364 sites were conserved in this dataset. The catalytic domain of this dataset was less conserved than in SSRV and the C-terminal somewhat more so. The third RCR motif was considerably more variable than in other datasets.



A **B**

Figure 4.7: A. WDV. **B.** USV conservation. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

The WDV set was conserved in 308 of 351 sites (Fig 4.7). WDV has a novel sequence TSPFS, slightly N-terminal of the third RCR motif, which is not present in the other mastreviruses. The C-terminal region and regions between the RCR motifs were again the most variable. 338 of 352 sites were conserved in the USV dataset, reflecting the low diversity of USV samples available. The catalytic domain and a region approximately 20 amino acids C-terminal of this domain were the most variable in this dataset.

Over all datasets several highly conserved sites were observed. The RCR motifs had several conserved sites. **L**x**Y** and **Y**xx**K** in the first and third RCR motif were conserved in all mastreviruses. **E**x**H** and **N**x**Q** situated in similar positions between the three RCR motifs were also completely conserved as was a region **EY**x**A** at the beginning of the oligomerisation domain. Several regions within or near the dNTP binding motifs were conserved including a particularly large block **NPKYGK** approximately 10 amino acids C-terminal of the final dNTP binding motif. Near the C-terminal of the protein **L**x**N**x**DE**x**W** was also conserved, and may be notable as it contains two acidic residues (Chapter 2 contains a detailed annotation of Rep domains and these conserved sites).

MSV-Kom and MSV-Set Comparison

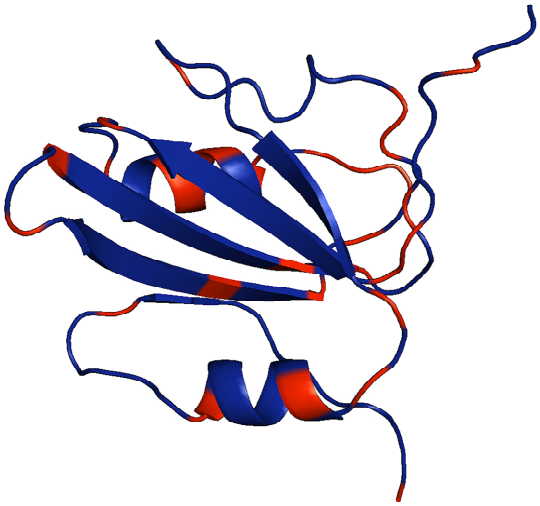


Figure 4.8: A. MSV-Kom and MSV-Set conservation. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

MSV-Kom and MSV-Set were conserved at 289 of 356 sites (Fig 4.8). Variable sites were predominantly in the catalytic domain, generally between RCR motifs. Within RCR motif one MSV-Kom possessed a lysine residue whereas MSV-Set possessed a Glutamine residue. The second and third RCR motifs had a similar disparity, of **LH** to **QC** in RCR two and **L** to **T** in RCR three. Sites 376-393 in the alignment (at the C-terminal end of the catalytic domain) were entirely non-conserved. Other active domains showed small variations. The myb-like transactivation domain had an **A** to **T**, the dNTP domain contained an **F** to **Y** and the C-terminal region contained an **RD** to **KE** mutation from MSV-Kom to Set.

Conservation by Host

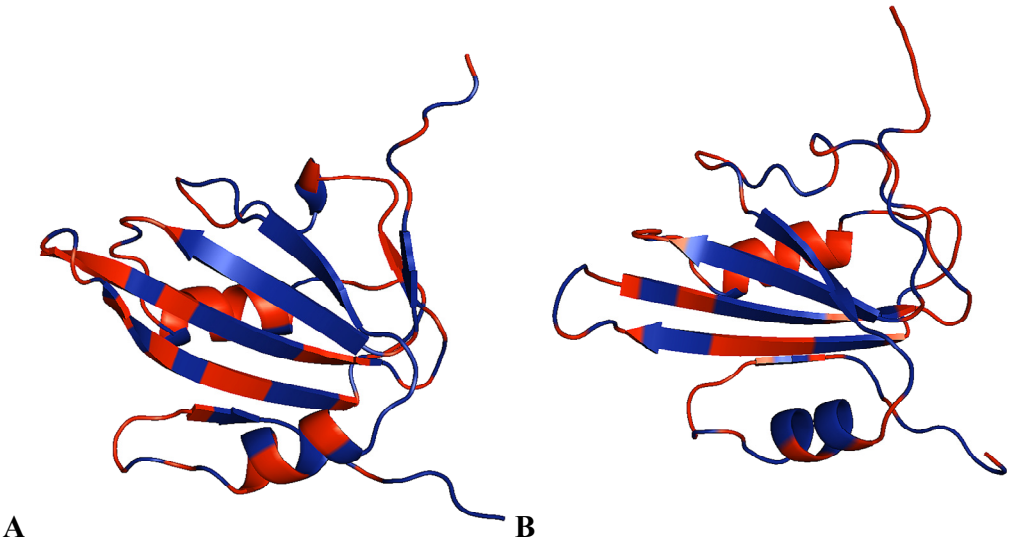


Figure 4.9: A. Digitaria conservation by host. **B.** Panicum conservation by host. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

The influence of host factors on mastreviruses evolution is as yet unknown, and may be an important factor. Rep in particular is thought to interact with several host genes, and so these effects may be particularly important. Sequences were subdivided according to the findings of Varsani *et al* (2008a). Digitaria infecting viruses were isolates of DSV and MSV-A4-B3, E-G and I. The panicum dataset is comprised of the isolates from PanSV and MSV-G.

177 of 366 sites were conserved in the digitaria dataset. RCR motifs were somewhat conserved within this dataset. A large tract of the region between RCR motifs two and three was conserved (centred on the **NIQ** site mentioned above). The oligomerisation domain was also highly conserved. Much of the dNTP binding motif and surrounding regions are variable. Notably the region between the first dNTP binding motif and the myb like transactivation domain are highly conserved.

192 of 370 sites were conserved in the panicum dataset. Again much of the RCR motif region was conserved. A region around the **NIQ** site was also conserved, however it

spanned fewer sites than in the digitaria set. The oligomerisation domain was slightly less conserved than in the digitaria set, as was the C-terminal domain of the protein.

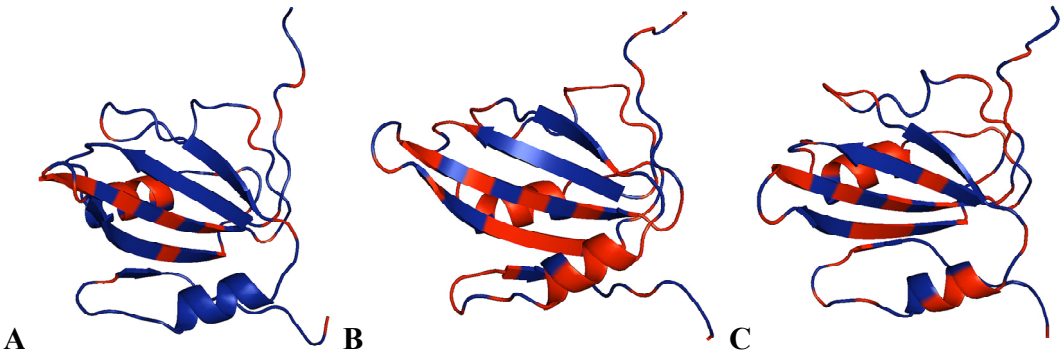


Figure 4.10: A. Setaria conservation by host. **B.** Sugarcane conservation by host. **C**. Urochloa conservation by host. Blue denotes conserved sites, red denotes variable sites. Sequences were aligned using CLUSTALW implemented in MEGA4 using a Gap opening penalty of 10 and a Gap extension penalty of 0.1. See the Supplementary material for the full alignment.

The setaria dataset was conserved of 290 of 357 sites. Variation was most prevalent in the catalytic domain and C-terminal regions of the protein, with a small central region near the C-terminal end of the catalytic domain being also variable. Within the catalytic region RCR sites were again somewhat conserved, with small conserved blocks around **NIQ** and a corresponding region between RCR motif one and two. Sugarcane was significantly more variable with only 190 of 367 sites conserved. Variation was distributed reasonably evenly along the protein. The dNTP binding motifs and surrounding regions were highly conserved. The RCR motifs and the **NIQ** region were again conserved in this set. The urochloa dataset was conserved at 230 of the 368 sites. Patterns were overall similar to those seen in the other two datasets except that regions around the dNTP binding motif and protein C-terminal were more conserved in urochloa.

.

4.4.1 Selection

To identify key regions under selection SLAC, FEL and IFEL selection analysis algorithms were used to identify sites under selection. FEL detected the highest number of sites under selection, followed by IFEL and then SLAC.

Selection was detectable in all datasets by one or more algorithms (Fig 4.11, 4.12). Sites under selection were identified throughout the replication protein, including numerous regions in the catalytic helicase domain. Purifying selection was predominant in these datasets. The MSV-A and PanSV datasets contained the most sites under detectable selection. MSV-B1 had more positive selection than the other MSV strains.The USV dataset had the lowest detectable selection, probably as a result of the low diversity between strains.

In the MSV-A-K sets, sites under purifying selection were found throughout the protein, including the catalytic domain. Sites under purifying selection were detected within the RCR motifs and around the conserved **NIQ** region between RCR motifs. Positive selection was detected in MSV-B1 and B-2. In MSV-B1 positive selection was found in the nucleotides coding for the R of the sequence **CAREAH** sequence which corresponds to a partly conserved region between RCR I and II. In MSV-B2 the site under selection was situated approximately 10 amino acids past the C-terminal of the catalytic region (Figure 4.11). The large MSV-A dataset also identified negative selection near the **NIQ** region, and also between the first two RCR motifs. Positive selection in this dataset was detected near the C-terminus of the catalytic region and around C-terminal regions of the protein. One site under selection was detectable just N-terminal of the Rbr binding motif.

Selection within the PanSV set was also found to be within functional domains A sites near the second RCR motif was under positive selection. A single site just C-terminal of the catalytic region was also under selection. A cluster of sites near the first dNTP binding motif were under detectable positive selection. Sites under purifying selection was again distributed throughout the protein, however several key sites were under selection. The conserved **NIQ** region and the region between RCR I and II were both under negative selection, as well as regions flanking the **L**x**C**x**E** motif. A large region from codons 294-318 (between the C-terminus and the second dNTP binding site) was highly conserved in this dataset and to a lesser extent in the SSV dataset.

Other datasets displayed conserved sites in similar patterns to those described above, however generally less selection was detected. C-terminal regions were under somewhat less detectable selection in the remaining sets. USV again had very little detectable selection.

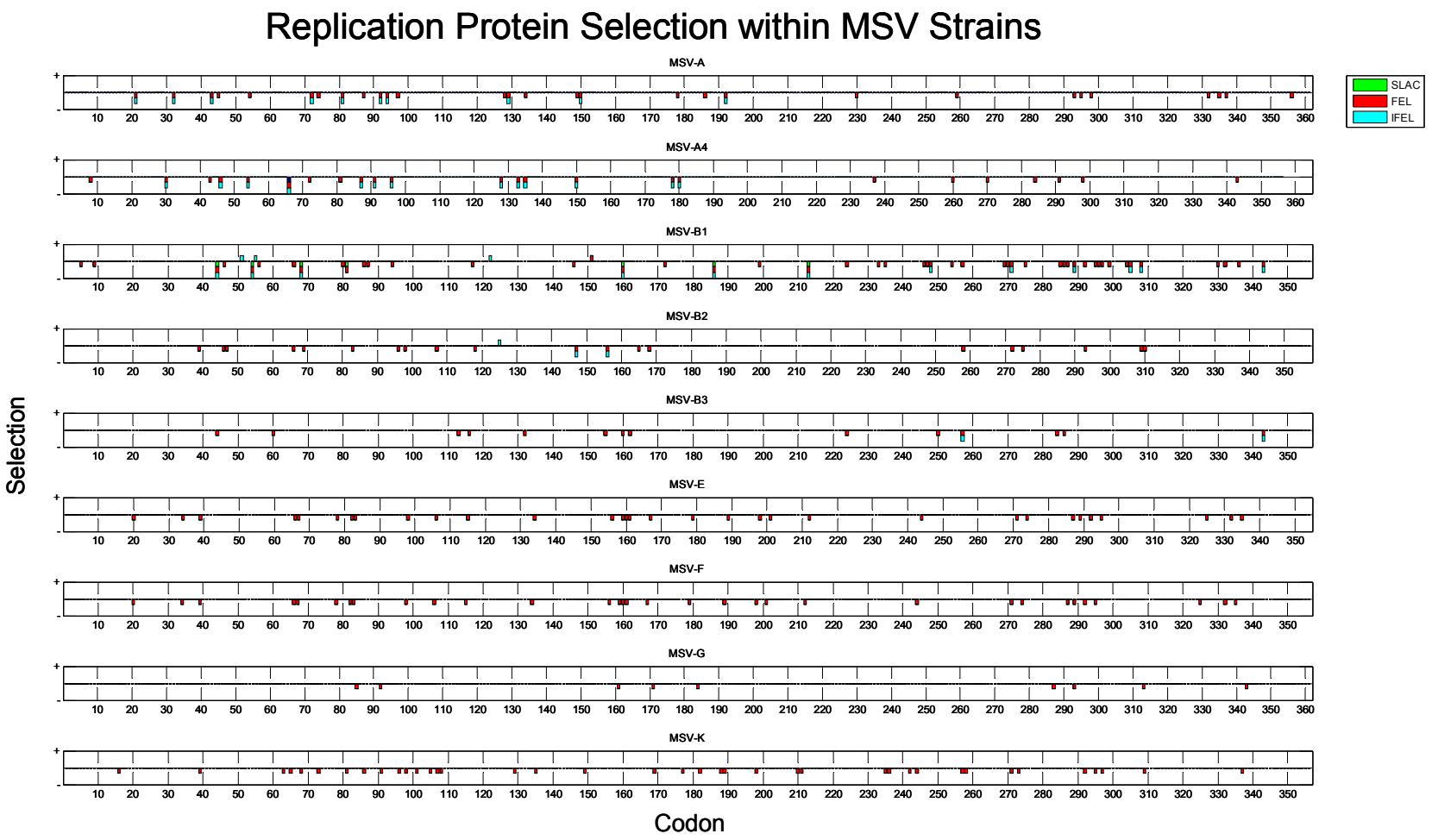


Figure 4.11: Selection detected within the MSV strains A-K. Colours denote which algorithm detected the selection. Green: SLAC Red: FEL Blue: IFEL Stacked columns indicate multiple detections by different algorithms. See supplementary material for full alignments.

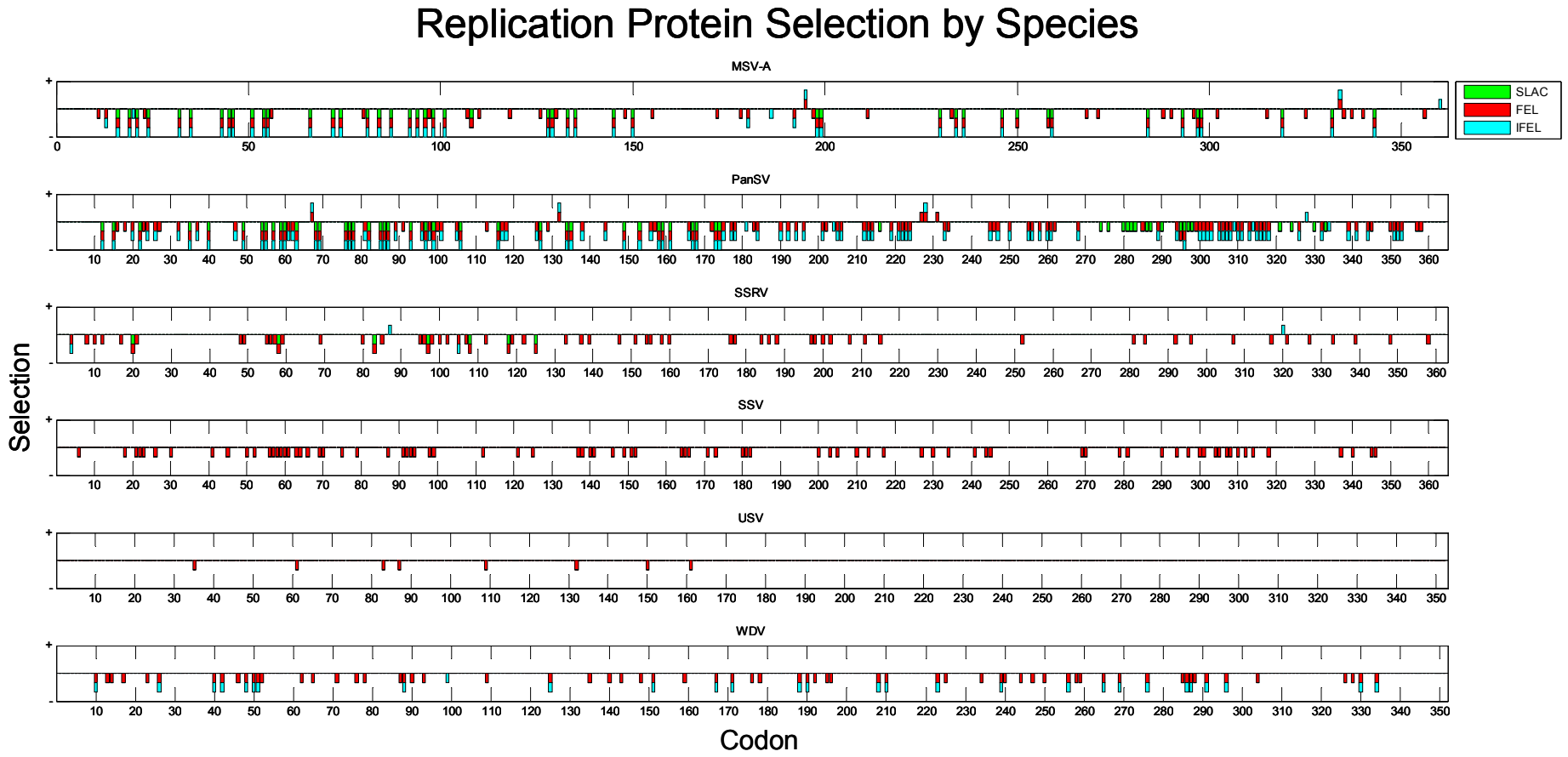


Figure 4.12: Selection detected over a range of mastrevirus species. Colours denote which algorithm detected the selection. Green: SLAC Red: FEL Blue: IFEL Stacked columns indicate multiple detections by different algorithms. See supplementary material for full alignments.

4.4.2 Selection by Host

In an attempt to detect links between host species and mastrevirus evolution, selection analysis by host plant was carried out. Detectable selection was found across the replication protein, including functional regions such as the catalytic domain. An increased number of sites under positive selection were found when compared to the 'by species' datasets except in the case of the panicum set. A number of these sites were detected by SLAC, which is the most conservative algorithm, making such detections likely to be true events. In all sets but panicum negative selection was detectable across the protein, and was relatively evenly distributed (Fig 4.14, 4.15). However positive selection within the panicum set was more predominant and concentrated mainly within two clusters, one within the catalytic domain near RCR III, the other situated near the dNTP binding domains.

3D modelling of these sites indicates that the random loop structures contain numerous sites under negative selection. Regions in the beta sheet under selection are situated close to RCR domains or the conserved regions between them. The alpha helix near the N-terminal region of the catalytic domain is also under detectable selection in numerous sets. The alpha helix contains the partially conserved **Y**xx**K** residues which are often under selection in these datasets. Interestingly the **Y**xx**K** positions nearby in space to a group of residues (**YILC**) on the beta sheet, which is conserved and under selection in a number of mastreviruses, particularly within MSV-A. This may have implications to the function of this region (Figure 4.13).

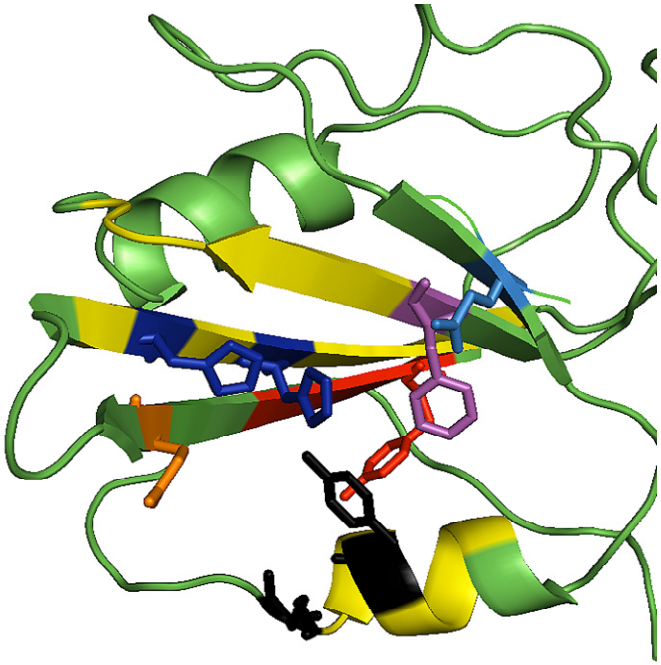


Figure 4.13: Key residues in the RCR domains of the catalytic region. Note how all of these residues position closely in 3D space. Key regions are coloured with likely functional amino acids showing. Yellow: RCR regions. Blue: Two histidine residues within the second RCR motif. Pink: Phenylalanine of the FLTY section of RCR I. Black: Highly conserved Y and K of the third RCR motif. Red: YILCARE region which is often partly conserved or under selection. Orange: Stick showing glutamic acid of the YILCARE region. Light Blue: Q of NIQ region

.

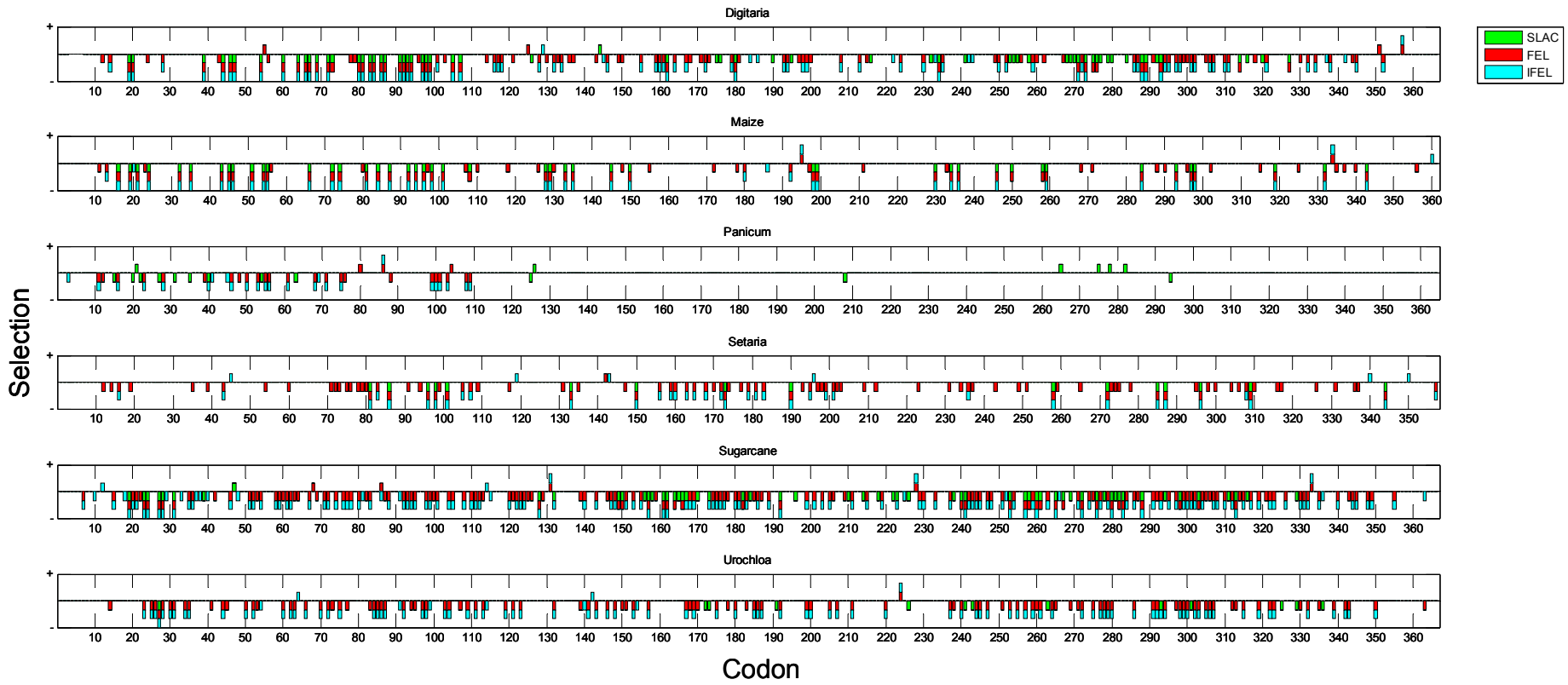


Figure 4.14: Selection detected over a range of mastrevirus species arranged by host plant. Colours denote which algorithm detected the selection. Green: SLAC Red: FEL Blue: IFEL Stacked columns indicate multiple detections by different algorithms. See supplementary material for full alignments.

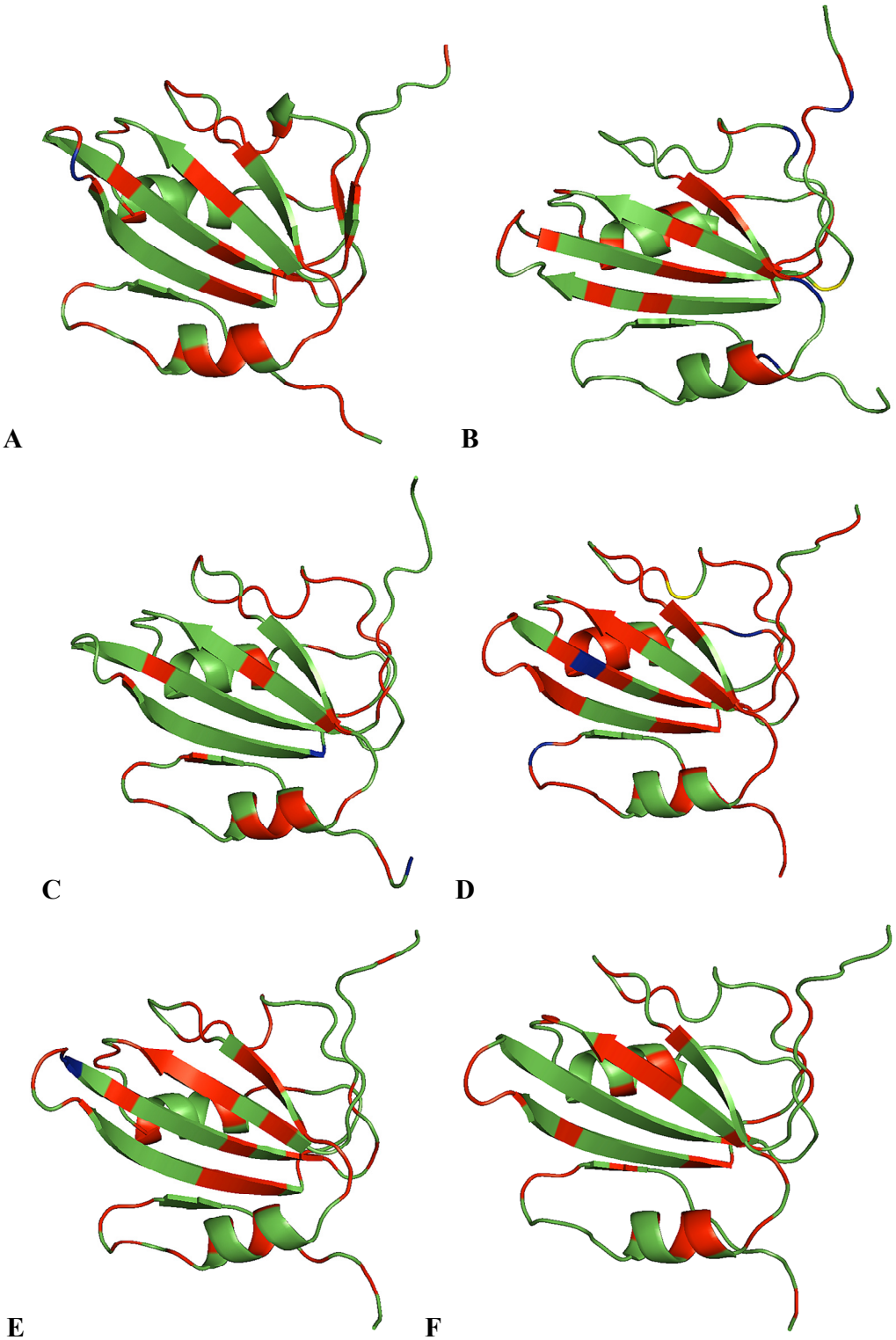


Figure 4.15: A. Digitaria by host selection **B**. Panicum by host selection. **C.** Setaria selection by host. **D.** Sugarcane selection by host **E.** Urochloa selection by host **F**. Maize selection by host. 3D models mapping selection within the catalytic region of these datasets. Red denotes negative selection; blue denotes positive selection and yellow denotes amino acids with nucleotides under both positive and negative selection.

4.4.3 Codon Usage

In order to determine if codon usage was a significant factor in mastrevirus evolution, and to compare Rep codon usage with usage in the coat protein, a codon usage analysis was carried out. A distinct preference for certain codons was observable (Table 1). Codon usage in this region varied distinctly 'by species' and was particularly pronounced in charged residues, e.g. AAG for lysine was strongly favoured in all datasets except digitaria and setaria in which usage of both codons was similar.

	Entire	Digitaria	Panicum	Setaria	Sugarcane	Maize	Urochloa		Entire	Digitaria	Panicum	Setaria	Sugarcane	Maize	Urochloa
UUU(F)	5.3(0.65)	5.2(0.63)	3.0(0.47)	3.8(0.49)	3.6(0.52)	6.6(0.79)	3.4(0.50)	**UCU(S)**	7.6(1.43)	8.2(1.56)	4.6(1.06)	8.2(1.56)	5.3(1.02)	8.3(1.38)	7.1(1.59)
UUC(F)	10.9(1.35)	11.3(1.37)	9.9(1.53)	11.8(1.51)	10.5(1.48)	10.2(1.21)	10.1(1.50)	**UCC(S)**	7.3(1.36)	6.6(1.26)	3.3(0.77)	7.4(1.41)	9.5(1.85)	8.6(1.42)	3.9(0.86)
UUA(L)	3.0(0.79)	4.2(1.12)	0.5(0.12)	3.6(0.99)	1.5(0.41)	3.0(0.79)	2.4(0.67)	**UCA(S)**	9.5(1.79)	9.8(1.87)	4.9(1.13)	8.2(1.56)	4.9(0.95)	12.9(2.13)	7.9(1.75)
UUG(L)	2.4(0.64)	2.8(0.75)	0.1(0.03)	2.6(0.72)	1.1(0.29)	3.1(0.83)	1.9(0.53)	**UCG(S)**	2.0(0.37)	1.7(0.33)	1.0(0.22)	2.0(0.38)	1.3(0.25)	2.3(0.38)	1.4(0.32)
CUU(L)	3.3(0.89)	3.5(0.94)	2.4(0.61)	2.6(0.72)	3.1(0.82)	3.6(0.95)	2.5(0.71)	**CCU(P)**	11.6(1.74)	13.3(1.88)	6.6(1.09)	13.8(1.99)	9.2(1.50)	12.8(1.86)	15.4(2.13)
CUC(L)	7.4(1.97)	6.4(1.73)	11.6(2.93)	7.0(1.93)	7.0(1.86)	6.3(1.67)	7.1(2.02)	**CCC(P)**	4.6(0.69)	3.2(0.46)	4.4(0.73)	3.4(0.49)	5.2(0.85)	5.6(0.82)	4.0(0.55)
CUA(L)	3.3(0.88)	2.8(0.74)	2.2(0.56)	3.6(0.99)	3.8(1.01)	3.6(0.95)	4.2(1.19)	**CCA(P)**	7.7(1.14)	9.3(1.31)	8.9(1.48)	8.8(1.27)	7.5(1.22)	5.8(0.84)	6.8(0.94)
CUG(L)	3.1(0.84)	2.7(0.72)	7.0(1.75)	2.4(0.66)	6.1(1.61)	3.0(0.81)	3.1(0.89)	**CCG(P)**	2.9(0.43)	2.4(0.34)	4.2(0.70)	1.8(0.26)	2.6(0.43)	3.3(0.49)	2.7(0.38)
AUU(I)	9.0(1.14)	9.8(1.21)	3.8(0.57)	7.6(0.95)	6.7(0.92)	11.8(1.41)	7.4(0.89)	**ACU(T)**	5.2(1.19)	5.4(1.41)	8.9(1.31)	5.6(1.45)	5.1(0.93)	3.9(1.18)	4.0(0.86)
AUC(I)	10.7(1.36)	11.4(1.42)	10.4(1.58)	12.6(1.57)	9.6(1.31)	9.5(1.13)	12.9(1.55)	**ACC(T)**	3.2(0.74)	1.4(0.36)	5.6(0.83)	1.4(0.36)	6.2(1.12)	2.7(0.80)	4.9(1.06)
AUA(I)	4.0(0.50)	2.9(0.36)	5.6(0.85)	3.8(0.47)	5.6(0.77)	3.8(0.46)	4.6(0.56)	**ACA(T)**	6.7(1.55)	6.2(1.62)	11.6(1.72)	5.8(1.51)	7.4(1.34)	4.8(1.43)	7.4(1.60)
AUG(M)	7.4(1.00)	8.9(1.00)	4.9(1.00)	8.8(1.00)	5.0(1.00)	7.5(1.00)	5.9(1.00)	**ACG(T)**	2.3(0.52)	2.3(0.60)	1.0(0.14)	2.6(0.68)	3.4(0.61)	2.0(0.60)	2.2(0.48)
GUU(V)	5.9(1.36)	6.2(1.47)	1.9(0.46)	6.6(1.53)	2.4(0.46)	6.0(1.46)	4.6(1.04)	**GCU(A)**	3.2(0.84)	1.6(0.47)	4.7(1.12)	1.4(0.44)	5.4(1.04)	2.3(0.64)	3.9(0.93)
GUC(V)	3.8(0.88)	3.4(0.81)	8.0(1.94)	4.0(0.93)	7.9(1.54)	2.1(0.50)	5.7(1.27)	**GCC(A)**	5.9(1.57)	5.5(1.61)	5.7(1.34)	6.2(1.94)	7.1(1.37)	6.1(1.69)	5.2(1.25)
GUA(V)	3.9(0.90)	4.1(0.99)	2.7(0.66)	4.2(0.98)	4.7(0.92)	4.9(1.19)	4.5(1.00)	**GCA(A)**	5.6(1.48)	6.4(1.86)	5.1(1.21)	5.0(1.56)	6.9(1.33)	6.0(1.66)	7.0(1.68)
GUG(V)	3.7(0.85)	3.0(0.73)	3.9(0.94)	2.4(0.56)	5.5(1.08)	3.5(0.85)	3.1(0.69)	**GCG(A)**	0.4(0.11)	0.2(0.07)	1.4(0.33)	0.2(0.06)	1.4(0.26)	0.0(0.00)	0.6(0.14)
UAU(Y)	6.3(0.76)	7.8(1.00)	2.8(0.26)	7.2(0.88)	2.8(0.31)	6.2(0.86)	5.8(0.68)	**UGU(C)**	5.3(1.35)	6.7(1.67)	2.6(0.62)	4.8(1.26)	2.3(0.76)	6.5(1.63)	5.2(1.30)
UAC(Y)	10.2(1.24)	7.8(1.00)	18.9(1.74)	9.2(1.12)	15.1(1.69)	8.3(1.14)	11.4(1.33)	**UGC(C)**	2.5(0.65)	1.3(0.33)	5.6(1.38)	2.8(0.74)	3.7(1.24)	1.5(0.37)	2.8(0.70)
UAA(*)	0.4(1.60)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	**UGA(*)**	0.3(1.38)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)
UAG(*)	0.0(0.02)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.0(0.00)	**UGG(W)**	9.7(1.00)	10.4(1.00)	9.6(1.00)	10.0(1.00)	10.7(1.00)	9.9(1.00)	11.3(1.00)
CAU(H)	5.5(1.17)	5.3(1.15)	4.9(0.94)	5.0(1.09)	5.6(1.02)	5.8(1.29)	4.8(0.89)	**CGU(R)**	2.8(1.01)	2.7(1.10)	1.6(0.47)	2.6(1.08)	1.3(0.48)	2.0(0.73)	1.1(0.39)
CAC(H)	3.9(0.83)	3.9(0.85)	5.5(1.06)	4.2(0.91)	5.4(0.98)	3.2(0.71)	6.0(1.11)	**CGC(R)**	0.8(0.28)	0.1(0.04)	2.2(0.63)	0.6(0.25)	2.9(1.09)	0.0(0.02)	1.1(0.42)
CAA(Q)	2.6(0.37)	2.4(0.32)	2.0(0.37)	3.8(0.46)	5.9(0.77)	1.8(0.24)	3.5(0.46)	**CGA(R)**	1.6(0.58)	1.3(0.55)	1.5(0.44)	1.6(0.67)	0.5(0.20)	1.9(0.68)	1.9(0.68)
CAG(Q)	11.8(1.63)	12.8(1.68)	8.9(1.63)	12.8(1.54)	9.4(1.23)	13.7(1.76)	11.9(1.54)	**CGG(R)**	1.0(0.36)	0.3(0.11)	5.2(1.49)	0.4(0.17)	1.3(0.48)	0.4(0.14)	1.7(0.63)
AAU(N)	8.9(0.94)	9.4(0.95)	5.4(0.77)	9.8(0.94)	7.1(0.70)	9.9(0.99)	8.3(0.95)	**AGU(S)**	2.2(0.41)	2.4(0.47)	1.9(0.45)	3.0(0.57)	3.9(0.76)	2.2(0.36)	0.8(0.18)
AAC(N)	10.2(1.06)	10.3(1.05)	8.7(1.23)	11.0(1.06)	13.2(1.30)	10.0(1.01)	9.1(1.05)	**AGC(S)**	3.4(0.63)	2.7(0.52)	10.3(2.38)	2.8(0.53)	6.0(1.16)	2.0(0.33)	5.9(1.31)
AAA(K)	11.4(0.82)	14.9(0.99)	5.6(0.40)	15.6(1.07)	9.5(0.75)	11.7(0.78)	10.8(0.78)	**AGA(R)**	5.5(1.99)	5.8(2.40)	5.2(1.49)	4.8(2.00)	5.5(2.08)	6.2(2.23)	7.1(2.63)
AAG(K)	16.6(1.18)	15.1(1.01)	21.9(1.60)	13.6(0.93)	15.8(1.25)	18.4(1.22)	16.7(1.22)	**AGG(R)**	4.9(1.78)	4.4(1.80)	5.1(1.48)	4.4(1.83)	4.5(1.67)	6.2(2.21)	3.4(1.24)
GAU(D)	12.8(1.34)	12.6(1.48)	5.7(0.50)	12.8(1.36)	8.1(0.89)	16.0(1.67)	11.7(1.24)	**GGU(G)**	2.7(0.88)	2.8(0.87)	3.8(1.19)	3.2(0.98)	3.6(1.05)	2.0(0.72)	3.0(0.99)
GAC(D)	6.3(0.66)	4.4(0.52)	17.3(1.50)	6.0(0.64)	10.1(1.11)	3.1(0.33)	7.1(0.76)	**GGC(G)**	2.1(0.67)	1.5(0.47)	3.2(1.00)	1.2(0.37)	3.2(0.92)	1.6(0.56)	2.5(0.83)
GAA(E)	11.8(0.93)	12.5(0.95)	8.6(0.69)	13.4(1.09)	9.1(0.81)	12.9(0.96)	8.6(0.69)	**GGA(G)**	3.6(1.16)	4.4(1.34)	3.9(1.23)	4.2(1.29)	3.7(1.07)	3.1(1.11)	2.5(0.83)
GAG(E)	13.8(1.07)	13.8(1.05)	16.4(1.31)	11.2(0.91)	13.4(1.19)	14.0(1.04)	16.4(1.31)	**GGG(G)**	4.0(1.29)	4.3(1.32)	1.9(0.58)	4.4(1.35)	3.4(0.97)	4.5(1.61)	4.1(1.35)

Table 1: Codon Usage in the Replication protein over the entire dataset and subdivided by host species. Red text denotes codons that show an observable usage bias.

4.5 Discussion

The mastrevirus Replication protein is the only absolute requirement for mastrevirus Replication (Gutierrez, 2000). As such it plays an important part in the life cycle of the virus. Prior studies have identified several key sequences on the Replication protein and have solved the structure of the catalytic domain. This study aimed to use the large number of sequences now available to identify further conserved sites and to detect regions on this protein under selection. This data would then be used to identify regions that are evolving under selective forces and to identify new domains that may be important to the function of Replication protein.

4.5.1 Sequence alignment and Conservation

This study made use of CLUSTAL W (Thompson *et al*., 1994) for pairwise and multiple sequence alignment. A detailed discussion of sequence alignment can be found in chapter 3. By using these alignments Replication protein coding regions, Rep and RepA, were identified in all species and spliced together to form the full length Replication protein. Functional sites that had previously been detected were identified by identity to known sequences. Rolling circle motifs (Koonin & Ilyina, 1992), the catalytic domain, the oligomerisation domain, the retinoblastoma binding motif (Arguello-Astorga *et al*., 2004), myb-like transactivation domain (Horvath et al., 1998) and dNTP binding motifs (Xie *et al*., 1999) were all identified using data from previous studies.

Conserved sites identified in this study confirm the importance of these genomic regions but indicate several nearby domains that may be important. Conservation within the RCR motifs and the dNTP binding motifs was high. Within the RCR regions the **Y**xx**K** residues, which had been determined to be vital to this region's activity, we found to be entirely conserved amongst all mastreviruses, providing further evidence for the importance of this region. RCR I contained two conserved residues **L** and **Y** within the **FLTY** region previously identified, suggesting that these two residues are crucial for the

region's activity. RCR III's **V**x**DY**xx**K** residue was also found to be highly conserved, with the **Y** and **K** being maintained throughout the mastreviruses.

Nearby these regions we identified two sites that may be relevant to this region's function. Between RCR I and II two residues **E**x**H** were conserved, within a greater region which had various degrees of conservation across different datasets. Between RCR II and III a **N**x**Q** sequence, often more conserved as **NIQ** was also observed (Fig 4.16). These regions were both positioned similar distances between each RCR motif which may suggest some key property of these regions, particularly given the acidic residue within the **E**x**H** region. A third region in this area, specific to WDV was identified. This sequence **TSPFS** was found slightly N-terminal of the NIQ region, and was completely conserved amongst WDV species. This may have some relevance to the different host plant of WDV.

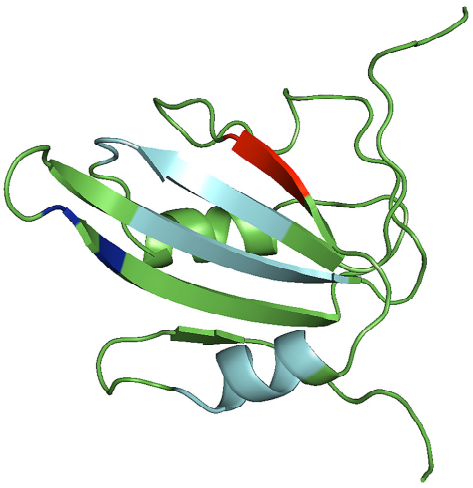


Figure 4.16: Locations of the NIQ (Red) and ExH (Blue) regions and surrounding sites. RCR regions are shown in cyan in MSV-A. These sites were often conserved, and their location indicates they may be important to the function of this region.

Two conserved regions near the dNTP motifs were identified. One was positioned between the two motifs, and had the sequence **HNY**. A second conserved region **NPKYGK** was identified near the C-terminal end of the protein. Given its continuous nature and proximity to the dNTP binding domains, this region potentially is important and may warrant further study.

Analysis of the MSV-Kom and MSV-Set region proved largely inconclusive, identifying numerous small residue changes that may be responsible for larger effects. It was notable that the favouring of Lysine in functional regions seemed to be still somewhat apparent in the Replication protein. Analysis by host was also inconclusive, but provided further evidence for conserved regions discussed above.

4.5.2 Selection

Selection analysis based on dN/dS estimation was carried out on all datasets using HyPhy implemented on the Datamonkey webserver. A detailed discussion of these procedures and their advantages and disadvantages is found in chapter 3. Using SLAC, FEL and IFEL algorithms numerous sites under selection were identified. Detectable selection was observable across the replication protein. Selection was predominantly negative, which was expected due to the numerous functional regions on the replication protein. Negative selection was detectable around several of the conserved regions identified in this study, providing a mechanism for maintenance of parts of these regions. Positive selection was generally associated with either the catalytic region, or a region surrounding the dNTP binding domain.

4.5.3 Selection by Host

Selection by host analysis resulted in increased detection of selection by all three algorithms. Selection was detected across the protein and was generally more predominant than in the 'by species' dataset . More positive selection was identified in these sets, particularly within the panicum group. This may indicate panicum has a particularly strong selective effect on certain regions of Rep. It was notable that relatively few sites under selection were detected outside the catalytic domain. This may indicate that this region is of increased significance in adaptation to a panicum host, or that this region is under more selection in mastreviruses in general. The relatively even distribution of negative selection across the remaining datasets suggests numerous sites contribute to the structure and function of Rep, making it difficult to pinpoint significant residues.

4.5.4 Codon Usage

Codon usage data suggests a preference for certain codons in the Rep protein. A similar preference for codons was seen in Rep as in coat protein. However the variation by host species is apparently more pronounced in the Replication protein, with variation of codon preference being apparent for certain codons. This is likely to provide a selection pressure on mastreviruses as they adapt to particular hosts.

4.5.5 Implications of Selection and Conservation Patterns

Data gathered in this study strongly support the importance of the initial catalytic region to the function of Rep. As this region is responsible for initiation of replication by nicking the TAATATTAC site this importance is hardly surprising. The strong negative selection operating across much of this region in all datasets suggests that there is less tolerance for change amongst key region; however these regions can often be deceptively small. The **Y**xx**K** of the RCR III motif is a good example, where only two sites seem to be absolutely conserved, yet provide the primary function of this region. This also serves to highlight the difficulty of functional analysis of mastrevirus protein domains.

The two conserved sites between the RCR motifs identified in this study were often subject to strong negative selection. This suggests further that these sites are likely to be important to the function or structure of this region. It is likely their positioning in the beta sheet near the two RCR regions is not coincidental, especially given the grouping of numerous other charged motifs in this region. The positioning of residues from all RCR motifs in this regions make it likely to be an active site within this domain, provided the 3D model accurately models the true position of these residues.

Positive selection detectable around the C-terminal regions, particularly in the 'by host' datasets may indicate that this region is significant to particular interactions with the host plant. The panicum dataset in particular shows several sites under strong selection, that were not detected in other datasets. These occur within a largely conserved region near the dNTP binding sites, suggesting host adaptation may be occurring within the panicum infecting species. C-terminal regions were under somewhat less negative selection in all datasets except sugarcane. This is interesting as GRAB proteins are thought to bind this

domain as part of a possible plant immune function that suppresses mastrevirus activity (Xie *et al*., 1999). A variable or randomly evolving Rep C-terminus may be advantageous to evading this mechanism. It is notable that C-terminal regions within the sugarcane dataset are under more negative selection. This may indicate specific importance of this region to sugarcane host adaptation, but could also indicate that variability in this region is not as advantageous to sugarcane infecting species. This would fit with the known susceptibility of sugarcane to a variety of mastrevirus species (Varsani *et al*., 2008a, Lawry *et al*., 2009).

Although selection was occurring across the protein, only approximately 50% or less of sites were under negative selection. This suggests neutral selection is a significant evolutionary force within the replication protein of mastreviruses. Highly conserved sites in the proteins are likely maintained through either an absolute requirement for functionality, which induces a sampling bias as only fit populations are likely to be sampled, or by random chance. Due to the high mutation rates seen in mastreviruses (between 10^{-3} and 10^{-5} substitutions per site per year) the chance of any site remaining conserved by chance over evolutionary time periods is slim (van der Walt *et al*., 2008). As Rep has numerous interactions with other gene regions, the interactions are likely to create an absolute requirement for certain domains, causing them to be conserved. Neutral selection is therefore the predominant force in mastrevirus evolution, however host adaptation, or purifying selection caused by selection pressures exerted by a host species can also play a part in the evolution of these viruses.

5 Conclusion

The aim of this study was to identify conserved regions and selection sites across functional domains in the mastrevirus genome. Mastrevirus functional domains are responsible for the virulence and replication of these viruses. An understanding of the key regions in these proteins and how they are evolving could prove crucial to predicting the behaviour of these viruses. As mastreviruses are a significant economic threat, particularly in developing countries, an in depth understanding of the mechanisms used by these viruses to function may be significant in the development of better techniques to mitigate their effects. The risk of spread of mastreviruses to new regions is significant and predicting factors responsible for adaptation to new hosts may allow prevention or reduction of this risk.

Mastreviruses are capable of rapid evolution into new host species. Adaptation to maize occurred over a period as short as 200 years (Harkins *et al*., 2009b). Although the maize virulent mastreviruses have been a primary focus for study, mastreviruses infect several other economically important crop species. Wheat, barley and sugarcane are known to be infected by mastreviruses, with as yet unknown economic impact. Sugarcane in particular appears to be highly susceptible to mastreviruses. A new sugarcane infecting mastrevirus called *Saccharum streak virus* (SacSV) was identified in this study. It shared a 66% identity to other mastreviruses and was closely related to *Urochloa streak virus.* Functional domains on this virus were identified by homology to known regions within other mastrevirus genomes. As relationships between sections of mastrevirus genomes are known to vary, analysis of individual genes was carried out. SacSV was found to be more closely related to *Sugarcane streak Egypt virus* within the movement protein. A general trend of low relatedness between movement protein and other gene regions was also observed, whilst coat and replication protein appeared to be more related. This corresponded well with earlier discoveries on gene networks with these regions (Martin *et al*., 2005).

Two hypotheses currently exist on mastrevirus evolution. One suggests codivergance with host species (Wu *et al*., 2008) and the other suggests neutral evolution to be the main process behind mastrevirus evolution (Harkins *et al*., 2009a). The predominant hypothesis is that mastreviruses are selectively neutral due to their high basal mutation rate (Harkins *et al*., 2009b), regions of the genome were tested for selection using bioinformatics tools. Samples from a large range of mastrevirus species were used to investigate regions under selection. The large array of samples also allowed high resolution conservation studies to be carried out, in order to identify new sequences, and further support previous evidence on key regions within mastrevirus genes. A recent study linking certain mastreviruses to a group of host species allowed analysis of links between host species and selection (Varsani *et al*., 2008a). These sites were mapped on homology modelled 3D structures of the coat protein and replication protein. This allowed more informative analysis of conserved regions and regions under selection.

Conserved sequences within the coat protein and movement protein were identified. These regions are thought to interact, however analysis revealed no obvious insights into these interactions. A few sites were conserved across the mastreviruses in the movement protein, which was expected due to the limited number of interactions it has with other regions. The transmembrane domain appeared to be highly conserved within species, probably due to the function of MP being inter-cellular transport. CP had numerous conserved sites, particularly towards the base of the N-terminal arm (Figure 3.9), which contains an active site, and in the C-terminal regions. Positioning of the arm may therefore be important to CP function. The function of the C-terminal region is more difficult to determine. Comparison of conserved sites by host yielded similar data. Comparison of maize virulent MSV-Kom with non-virulent MSV-Set identified several regions within MP and CP that may be important but also highlights the importance of numerous small residue changes to virulence within different species.

Sites under selection with CP and MP were found in numerous regions across the proteins. Detectable selection was predominantly negative, as expected within functional protein regions, however under 50% of sites were selected in each dataset. Several

regions within the DNA binding domain were under negative selection. Positive selection in the maize dataset NLS indicates that this domain may be significant to host adaptation. MP had concentrations of negative selection in the C-terminal region and the transmembrane domain, confirming the significance of the transmembrane domain and indicating the C-terminal region may be significant to MP function.

Within the replication protein numerous motifs with homology to previously identified regions in other mastreviruses were found. The genus wide importance of the RCR motifs was confirmed by highly conserved sites in each of these regions, and conserved regions in the vicinity of these were identified. NxQ and ExH residues are likely to have some significance to this region, due to both their conserved nature and the positions they map to on 3D structures of the catalytic domain. Similarly the dNTP binding motifs and oligomerisation domains were found to be more highly conserved than other sites within the replication protein.

Selection in Rep was spread across the entire protein, except notably in the panicum host dataset. Selection by host in the replication protein appeared to be highly significant, with more sites being detected in selection by host datasets. Sugarcane infecting species in particular appeared to be undergoing intense selection across the protein. Panicum infecting species were somewhat of an outlier, with fewer negative sites, and numerous sites under positive selection. This may suggest host adaptation within this region is more significant than in others. Both conserved and selected regions often mapped to several key sites within the catalytic domain and from this data a potential binding region was identified, containing numerous charged residues. Although Rep is known to interact with other proteins, none of the sites responsible for this could be identified in this study.

Conservation analysis has revealed several sites within mastreviruses that are likely to be vital to function. Selection on several of these sites was detectable, giving an obvious force for maintenance of these regions. Other conserved sites were under no detectable selection, indicating absolute conservation of these regions within samples. These regions are therefore vital across the genus. Several of these sites were used to detect likely active

regions on proteins in conjunction with 3D homology modelling techniques, constructing a basis for future studies *in vitro*. Unfortunately even given the known interactions between proteins, no obvious sites were detected in this study, highlighting the difficulties of genetic analysis of modular genomes.

Selection analysis was more successful at determining regions of likely import, identifying several potentially significant regions in all three proteins. Combinatorial analysis using 3D modelling of selected sites allowed more specific determination of the link between selected sites and likely functional or structural relationships, and often provided support for the importance of conserved or semi-conserved regions. Most significant however was the generally low level of negative or purifying selection detected. Only approximately 50% or less of sites were found to be under negative selection across all three proteins.

Relatively few sites were absolutely conserved or negatively selected in these ORFs and their functional domains. Functional regions are more likely to be under negative selection than either of the mastrevirus intergenic regions, and comprise a greater part of the total genome area. As the co-divergance hypothesis requires over 99% negative selection to be valid, this study provides strong evidence that this is not occurring within mastreviruses. A neutral hypothesis of mastrevirus evolution best fits this data and aligns with data gathered in earlier field studies.

Supplementary Material

Note: To view the supplementary material the user will require Pymol, available free at www.pymol.org.

6 References

Abascal F, Zardoya R, Posada D (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics,* **21**, 2104-2105

Ammar E, Gargani D, Lett JM, Peterschmitt M (2009). Large accumulations of maize streak virus in the filter chamber and midgut cells of the leafhopper vector *Cicaduina mbila*. *Arch Virol*,**154**, 255-262

Andersen MT, Richardson KA, Harbison SA, Morris BAM (1988). Nucleotide sequence of the geminivirus chloris striate mosaic virus. *Virology*, **164**, 443-449

Arguello-Astora G, Lopez-Ochoa L, Kong L, Orozco BM, Settlage SB, Hanley-Bowdoin L (2004). A novel motif in Geminivirus Replication proteins interacts with the plant retinoblastoma-related protein. *J. Virol.*. **78**. 4817-4826

Ball ED (1909). The leafhoppers of the sugar beet and their relation to the "curly leaf" condition. *U.S Dept Agric. Bur. Entomol.* **66**, 33

Bigarre´ L, Salah M, Granier M, Frutos R, Thouvenel J-C, Peterschmitt M (1999). Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Arch Virol.* **144**, 2331–2344

Bosque-Perez NA, Olojede SO, Buddenhagen IW (1998). Effect of maize streak virus disease on the growth and yield of maize as influenced by varietal resistance levels and plant stage at time of challenge. *Euphytica* **101**, 307-317

Bosque-Perez (2000). Eight decades of *maize streak virus* research. *Virus Res.* **71**, 107-121.

Boulton MI, Steinkellner H, Donson J, Markham PG, King DI, Davies JW (1989). Mutational Analysis of the Virion-sense Genes of Maize Streak Virus. *J. gen. Virol*. **70**, 2309-2323

Briddon RW, Pinner MS, Stanley J, Markham PG (1990). Geminivirus coat protein gene replacement alters insect specificity. *Virology* **177**, 85-94

Briddon RW, Lunness P, Chamberlin LC, Pinnern MS, Brundish H, Markham PG (1992). The nucleotide sequence of an infectious insect-transmissible clone of the geminivirus Panicum streak virus. *J. Gen. Virol.* **73**, 1041–1047

Campos-Olivas R, Louis JM, Clerot D, Gronenborn, B, Gronenborn AM. (2002). The structure of a Replication initiator unites diverse aspects of nucleic acid metabolism. *PNAS*. **99**, 10310-10315

Chatani M, Matsumoto Y, Mizuta H, Ikegami M, Boulton MI, Davies JW (1991). The nucleotide and genome structure of the geminivirus miscanthus streak virus. *J. Gen. Virol.* **72**, 2325-2331

Delano WL. "The PyMOL molecular graphics system." ***Delano Scientific LLC*****, San Carlos, CA, USA. http://www.pymol.org**

Dickinson VJ, Halder J, Woolston CJ (1996). The product of Maize Streak Virus ORF V1 is Associated with Secondary Plasmdesmata and Is First Detected with the Onset of Viral Lesions. *Virolog.* **220**, 51-59

Donson J, Accotto GP, Boulton MI, Mullineaux PM, Davies JW (1987). The nucleotide sequence of a geminivirus from Digitaria sanguinalis. *Virology* **161**, 160-169

Doyle JJ, Doyle JL (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* **19**, 11–154.

Duffy S, Holmes EC (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol* **82,** 957-965.

Duffy S, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet.* **9,** 267-276.

Ermolaeva MD (2001). Synonymous codon usage in Bacteria, *Curr. Issues. Mol. Biol.* **3**, 91-97

Eswar N, Marti-Renom MA, Webb B, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A (2006). Comparative Protein Structure Modeling With MODELLER. *Curr Protoc Bioinformatics*, John Wiley & Sons, Inc., Supplement **15**

Fauquet C, Stanley J (2003). 'Geminivirus classification and nomenclature: progress and problems', *Annals of Applied Biology* **142**, 165–189.

Fauquet CM, Stanley J (2005). Revising the way we conceive and name viruses below the species level: a review of geminivirus taxonomy calls for new standardized isolate descriptors. *Arch Virol* **150**, 2151–2179.

Fenoll C, Black DM, Howell SH (1988). The intergenic region of maize streak virus contains promoter elements involved in rightward transcription of the viral genome. *EMBO J* **7**, 1589-1596

Fenoll C, Schwarz JJ, Black DM, Schneider M, Howell SH (1990). The intergenic region of maize streak virus contains a GC-rich element that activates rightward transcription and binds maize nuclear factors. *Plant Mol Biol* **15**, 865-877

Fuller C (1901). Mealie variegation. In: 1st Report of the Government Entomologist, Natal, 1899–1900. Pietermaritzburg, Natal, South Africa: P. Davis & Sons, Government Printers.

Galvez GE, Castano M (1976). Purification of the whitefly-transmitted bean golden mosaic virus, *Turrialba* **26**. 205-207

Giddings NJ, Bennett CW, Harrison AL (1951). A tomato disease resembling curly top. *Phytopathology* **41,** 415-417

Gorbalenya AE, Koonin EV (1989). Viral proteins containing the purine NTP-binding sequence pattern. *Nucl. Acids Res.* **17,** 8413–8440.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**:696–704

Gutierrez C (2000). Geminiviruses and the plant cell cycle, *Plant Mol. Biol.* **43,** 763-772

Halley-Stott RP, Tanzer F, Martin DP, Rybicki EP (2007). The complete nucleotide sequence of a mild strain of Bean yellow dwarf virus. *Arch. Virol.* **152**, 1237-1240

Harkins GW, Delport W, Duffy S, Wood N, Monjane AL, Owor BE, Donaldson L, Saumtally S, Triton G, Briddon RW, Shepherd DN, Rybicki EP, Martin DP, Varsani A (2009a). Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virol J* (In Press)

Harkins GW, Martin, DP, Siobain D, Monjane AL, Shepherd DN, Windram OP, Owor BE, Donaldson L, van Antwerpen T, Sayed RA, Flett B, Ramusi M, Rybicki EP, Peterschmitt M, Varsani A (2009b). Dating the origins of the maize-adapted strain of maize streak virus, MSV-A, *J. Gen. Virol.* **90**, 3066-3074

Harrison, B.D., Barker, H., Bock, K.R., Guthrie, E.J., Meredith, G, Atkinson, M. (1977). Plant Viruses with circular single-stranded DNA. *Nature* **270**, 760-762

Hatta T, Francki RIB. (1978). The fine structure of chloris striate mosaic virus. *Virology* **92**, 428-435

Heyraud F, Matzeita AV, Schaefera S, Schella J, Gronenborn B (1993). The conserved nonanucleotide motif of the geminivirus stem-loop sequence promotes replicational release of virus molecules from redundant copies. *Biochimie* **75**, 605-615

Horvath GV, Pettko-Szandtner A, Nikovics K, Bilgin M, Boulton M, Davies JW, Gutierrez C, Dudits D (1998). Prediction of functional regions of the maize streak virus replication-associated proteins by protein-protein interaction analysis. *Plant Mol. Biol.* **38,** 699–712.

Hou Y, Sanders R, Ursin, VM, Gilbertson RL (2000) Transgenic Plants Expressing Geminivirus Movement Proteins: Abnormal Phenotypes and Delayed Infection by *Tomato mottle virus* in Transgenic Tomatoes Expressing the *Bean dwarf mosaic virus* BV1 or BC1 proteins. *MPMP* **13**, 297-308

Howarth AJ, Caton J, Bossert M, Goodman RM (1985). Nucleotide sequence of bean golden mosaic virus and a model for gene regulation in geminiviruses. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 3572-3576.

Howell SH (1985). Physical structure and genetic organisation of the genome of maize streak virus (Kenyan isolate). *Nucleic Acids Res.* **12**, 7359-7375.

Hughes FL, Rybicki EP, Kirby R (1993). Complete nucleotide sequence of sugarcane streak Monogeminivirus. *Arch Virol* **132**, 171–182

Ilyina TV, Koonin EV (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucl. Acids Res.* **20,** 3279–3285.

Köklü G, Ramsell JNE, Kvarnheden A (2007) The complete genome sequence for a Turkish isolate of *Wheat dwarf virus* (WDV) from barley confirms the presence of two distinct WDV strains. *Virus Genes,* **34**, 359-366

Kotlizky G, Boulton MI, Pitaksutheepong C, Davies JW, Epel BL (2000). Intracellular and Intercellular Movement of Maize Streak Geminivirus V1 and V2 Proteins Transiently Expressed as Green Fluorescent Protein Fusions. *Virology* **274**, 32-38

Kryazhimskiy S, Plotkin JB (2008). The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304

Lawry R, Martin DP, Shepherd DN, van Antwerpen T, Varsani A (2009). A novel sugarcane-infecting mastrevirus from South Africa. *Arch. Virol.* **154**, 1699-703

Lazarowitz SG (1987). The molecular characterization of geminiviruses. *Plant Mol. Biol. Reptr.* **4**, 177-192.

Lazarowitz SG, Pinder AJ, Damsteegt VD, Rogers SG (1989). Maize streak virus genes essential for systemic spread and symptom development. *EMBO J.* **8**, 1023– 1032.

Lefeuvre P, Lett J, Reynaud B, Martin DP (2007a). Avoidance of protein fold disruption in natural virus recombinants. *PLoS Pathog* **3,** e181.

Lefeuvre P, Martin DP, Hoareau M, Naze F, Delatte H, Thierry M, Varsani A, Becker N, Reynaud B, Lett, JM (2007b). Begomovirus 'melting pot' in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *J Gen Virol* **88**, 3458–3468.

Lefeurve P, Lett J-M, Varsani A, Martin DP (2009). Widely conserved recombination patterns amongst single stranded DNA viruses. *J Virol*. **83**, 2697-707

Liu H, Boulton KJ, Davies JW (1997a). Maize streak virus coat protein binds single- and double-stranded DNA *in vitro*. *J. Gen. Virol.* **78**, 1265-1270

Liu L, van Tonder T, Pietersen G, Davies JW, Stanley J (1997b). Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *J. Gen. Virol.* **78**, 2113-2117

Liu H, Boulton MI, Thomas CL, Prior DAM, Oparka KJ, Davies JW (1999). Maize Streak Virus Coat Protein is Karyophillic and Facilitates Nuclear Transport of Viral DNA. *MPMP* **12**, 894-900

Liu H, Boulton MI, Oparka KJ, Davies JW (2001). Interaction of the movement and coat proteins of *Maize streak virus*: implications for the transport of viral DNA. *J. Gen. Virol.* **82**, 35-44

MacDowell SW, Macdonald H, Hamilton WD, Coutts RH, Buck KW (1985). The nucleotide sequence of cloned wheat dwarf virus DNA. *EMBO J.* **4**, 2173, 2180

Makkouk KM, Dafalla G, Hussein M, Kumari SG (1995). The natural occurrence of chickpea chlorotic dwarf geminivirus in chickpea and faba bean in Sudan. *J Phytopathol* 143, 465–466

Martin DP, Willment JA, Rybicki EP (1999). Evaluation of maize streak virus pathogenicity in differentially resistant Zea mays genotypes. *Phytopathology* **89,** 695–700.

Martin DP, Willment JA, Billharz R, Velders R, Odhiambo B, Njuguna J, James D, Rybicki EP (2001). Sequence Diversity and Virulence in *Zea mays* of *Maize Streak Virus* Isolates. *Virology* **288**, 247-255

Martin DP, van Der Walt E, Posada D, Rybicki EP. (2005). The Evolutionary Value of Recombination is Constrained by Genome Modularity. *PLoS Genet.* **1**, e51

Martin DP (2009). Recombination detection and analysis using RDP3. *Methods Mol Biol.* **537**, 185-205.

Monfreda C, Ramankutty N, Foley JA (2008). Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical CY* **22**, GB1022

Morris-Krsinich BAM, Mullineaux PM, Donson J, Boulton MI, Markham PG, Short MN, Davies JW (1985). Bidirectional transcription of maize streak virus DNA and identification of the coat protein gene. *Nucleic Acids Res.* **13**, 7237–7256.

Morris BAM, Richardson KA, Haley A, Zhan X, Thomas JE (1992). The nucleotide sequence of the infectious cloned DNA component of tobacco yellow dwarf virus reveals features of geminiviruses infecting monocotyledonous plants. *Virology,* **187**, 633-642

Nahid N, Amin I, Mansoor S, Rybicki EP, van der Walt E, Briddon RW (2008). Two dicot-infecting mastreviruses (family Geminiviridae) occur in Pakistan.*Arch Virol.* 153(8),1441-51.

Oluwafemi S, Varsani A, Monjane AL, Shepherd DN, Owor BE, Rybicki EP, Martin DP (2008). A new African streak virus species from Nigeria, *Arch Virol*,**153**, 1407-1410

Owor BE, Shepherd DN, Taylor NJ, Edema R, Monjane AL, Thomson JA, Martin DP, Varsani A (2007a). Successful application of FTA classic card technology and use of bacteriophage phi29 DNA polymerase for large-scale field sampling and cloning of complete maize streak virus genomes. *J Virol Methods* **140**, 100–105

Owor BE, Martin DP, Shepherd DN, Edema R, Monjane AL, Rybicki EP, Thomson JA, Varsani A (2007b). Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. *J Gen Virol* **88**,3154–3165

Peterschmitt M, Reynaud B, Sommermeyer G, Baudin P (1991). Characterization of maize streak virus isolates using monoclonal and polyclonal antibodies and by transmission to a few hosts. *Plant Dis* **75**, 27–32

Pinner MS, Markham PG (1990). Serotyping and Strain Identification of Maize Streak Virus Isolates, *J Gen Virol*, **71**, 1635-1640

Pond SLK, Frost SDW, Muse SV (2005a). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679

Pond SLK, Frost SDW (2005b). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments *Bioinformatics*, **21,** 2531-2533

Pond SLK, Frost SDW (2005c). Not so different after all: A comparison of methods for detecting amino acid sites under selection, *Mol. Biol. & Evol.* **22**, 1208-1222

Pond SLK, Frost SDW, Grossman Z, Gravenor MB, Richman DD, Brown AJL. (2006). Adaptation to different human populations by HIV-1 revealed by codon based analysis. *PLoS Comp. Biol.* **2**, e62

Posada D, Buckley TR. (2004). Model selection and model averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over Likelihood ratio tests. *Syst. Biol* **53***,* 793-808

Posada D (2006). ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research* **37**: 700-703

Saunders K, Bedford ID, Yahara T, Stanley J (2003). The earliest recorded plant virus disease. *Nature* **422**, 831

Schubert J, Habeku A, Kazmaier K, Jeske H (2007). Surveying cereal-infecting geminiviruses in Germany-diagnostics and direct sequencing using rolling circle amplification. *Virus Res* **127**, 61-70

Shepherd DN, Martin DP, McGivern DR, Boulton MI, Thomson JA, Rybicki EP (2005). A three-nucleotide mutation altering the *Maize streak virus* Rep pRBR-interaction motif reduces symptom severity in maize and partly reverts at high frequency without restoring pRBR-Rep binding. *J. Gen. Virol.* **86**, 803-813

Shepherd DN, Martin DP, van der Walt E, Dent K, Varsani A & Rybicki, E.P. (2010). Maize streak virus: an old and complex 'emerging' pathogen. *Mol. Plant. Pathol.* **11**, 1-12

Shepherd DN, Varsani A, Windram OP, Lefeuvre P, Monjane AL, Owor BE, Martin PE (2008). Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and La Reunion, *Arch Virology*, **153**, 605-609

Stanley J, Markham PG, Callis RJ, Pinner MS (1986). The nucleotide sequence of an infectious clone of the geminivirus beet curly top virus. *EMBO Journal* **5**, 1761-1767.

Stanley J, Bisaro DM, Briddon RW, Brown JK, Fauquet CM, Harrison BD, Rybicki EP, Stenger DC (2005). Geminiviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds) Virus Taxonomy (VIIIth Report of the ICTV). Elsevier/Academic Press, London, pp 301-306

Storey HH (1925). The transmission of streak disease of maize by the leafhopper Balclutha mbila Naude´. *Ann. Appl. Biol.* **12**, 422–439.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**,1596–1599

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673–4680

van Antwerpen T, McFarlane SS, Buchanan GF, Shepherd DN, Martin DP, Rybicki EP, Varsani A (2008). First report of Maize streak virus field infection of sugarcane in South Africa. *Plant Dis* **92**, 982.

van der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP (2008). Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virol J*. **5**, 104.

van der Walt E, Rybicki EP, Varsani A, Polston JE, Billharz R, Donaldson L, Monjane AL & Martin DP.(2009). Rapid host adaptation by extensive recombination.*J Gen Virol*. **90**:734-46

Varsani A, Shepherd DN, Monjane AL, Owor BE, Erdmann JB, Rybicki EP, Peterschmitt M, Briddon RW, Markham PG, Oluwafemi S, Windram OP, Lefeuvre P, Lett J-M, Martin DP (2008a). Recombination, decreased host specificity and increased motility may have driven the emergence of Maize streak virus as an agricultural pathogen. *J Gen Virol*. **89**, 2063 – 2074

Varsani A, Oluwafemi S, Windram OP, Shepherd DN, Monjane AL, Owor BE, Rybicki EP, Lefewvre P, Martin DP (2008b). Panicum streak virus diversity is similar to that observed for maize streak virus, *Arch Virol*, **153**: 601-604

Varsani A, Shepherd DN, Dent K, Monjane AL, Rybicki EP, Martin DP (2009). A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virol. J.* **6**, 36

Woolston CJ, Barker R, Gunn H, Boulton MI, Mullineaux PM (1988). Agroinfection and nucleotide sequence of cloned wheat dwarf virus DNA. *Plant Mol Biol* **11**, 35-43

Wu B, Melcher U, Guo X, Wang X, Fan L, Zhou G (2008). Assessment of codivergence of Mastreviruses with their plant hosts. *BMC Evol. Biol.* **8,** 335

Xie Q, Suárez-López P, Gutiérrez C (1995). Identification and analysis of a retinoblastoma binding motif in the replication protein of a plant DNA virus: requirement for efficient viral DNA replication. *EMBO J.* **14,** 4073–4082.

Xie Q, Sanz-Burgos A.P, Guo H, García JA, Gutierrez C (1999). GRAB proteins, novel members of the NAC domain family, isolated by their interaction with a geminivirus protein. *Plant Mol. Biol*. **39**, 647–656.

Xie J, Wang X, Liu Y, Peng, Zhou G (2007). First report of the occurrence of Wheat dwarf virus in wheat in China. *Plant Dis* **91**, 111

Yazdi HR, Heydarnejad J, Massumi H (2008). Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel nonanucleotide. *Virus Genes* **36**, 539-45

Zhang W, Olson NH, Baker TS, Faulkner L, Agbandje-Mckenna M, Boulton MI, Davies JW, Mckenna R (2001). Structure of the Maize Streak Virus Geminate Particle. *Virology* **279**, 471-477